6

Cloning strategies

In the previous two chapters I examined the two essential components of genetic engineering, these being (i) the ability to cut, modify and join DNA molecules *in vitro*, and (ii) the host/vector systems that allow recombinant DNA molecules to be propagated. With these components at his or her disposal, the genetic engineer has to devise a cloning strategy that will enable efficient use of the technology to achieve the aims of the experiment. In Chapter 1 I showed that there are basically four stages to any cloning experiment (Fig. 1.1), involving **generation** of DNA fragments, **joining** to a vector, **propagation** in a host cell, and **selection** of the required sequence. In this chapter I examine some of the strategies that are available for completing the first three of these stages by the traditional methods of gene cloning, largely restricting the discussion to cloning eukaryotic DNA in *E. coli*. The use of the **polymerase chain reaction** (PCR) in amplification and cloning of sequences is discussed in Chapter 7, as this is now a widely used protocol which in some cases bypasses standard cloning techniques. Selection of cloned sequences is discussed in Chapter 8, although the type of selection method that will be used does have to be considered when choosing host/vector combinations for a particular cloning exercise.

6.1 Which approach is best?

The complexity of any cloning experiment depends largely on two factors: (i) the overall aims of the work, and (ii) the type of source material from which the nucleic acids will be isolated for cloning. Thus a strategy to isolate and

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

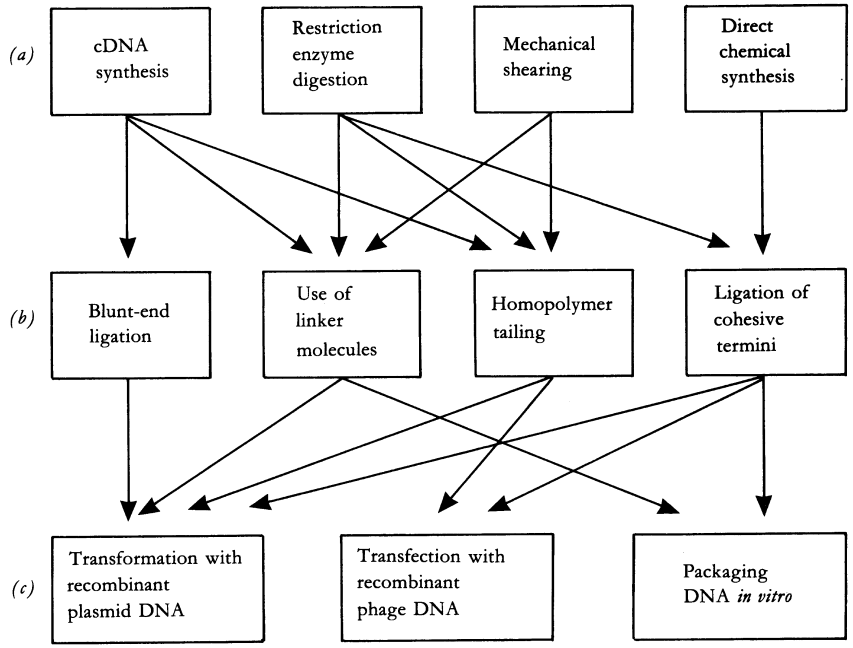


Fig. 6.1. Routes available for cloning. The possibilities for (*a*) the generation of DNA fragments, (*b*) joining to a vector and (*c*) introducing the recombinant DNA into a host cell. Preferred routes are indicated by arrows. Redrawn from Old and Primrose (1989), *Principles of Gene Manipulation*, Blackwell. Reproduced with permission.

sequence a relatively small DNA fragment from *E. coli* will be different (and will probably involve fewer stages) from a strategy to produce a recombinant protein in a transgenic eukaryotic organism. There is no single cloning strategy that will cover all requirements. Each project will therefore be unique, and will present its own set of problems that have to be addressed by choosing the appropriate path through the maze of possibilities (see Fig. 6.1). Fortunately, most of the confusion can be eliminated by careful design of experiments and rigorous interpretation of results.

When dealing with eukaryotic organisms, the first major decision is whether to begin with messenger RNA (mRNA) or genomic DNA. Although the DNA represents the complete genome of the organism, it may contain non-coding DNA such as introns, control regions and repetitive sequences. This can sometimes present problems, particularly if the genome is large and the aim is to isolate a single-copy gene. However, if the primary interest is in the control of gene expression, it is obviously necessary to isolate the control sequences, so genomic DNA is the only alternative.

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

Messenger RNA has two advantages over genomic DNA as a source material. Firstly, it represents the genetic information that is being expressed by the particular cell type from which it is prepared. This can be a very powerful preliminary selection mechanism, as not all the genomic DNA will be represented in the mRNA population. Also, if the gene of interest is highly expressed, this may be reflected in the abundance of its mRNA, and this can make isolation of the clones easier. A second advantage of mRNA is that it, by definition, represents the coding sequence of the gene, with any introns having been removed during RNA processing. Thus production of recombinant protein is much more straightforward if a clone of the mRNA is available.

Although genomic DNA and mRNA are the two main sources of nucleic acid molecules for cloning, it is possible to synthesise DNA *in vitro* if the amino acid sequence of the protein is known. Whilst this is a laborious task for long stretches of DNA, it is a useful technique in some cases, particularly if only short sections of a gene need to be synthesised to complete a sequence prior to cloning.

Having decided on the source material, the next step is to choose the type of host/vector system. Even when cloning in *E. coli* hosts there is still a wide range of strains available, and care must be taken to ensure that the optimum host/vector combination is chosen. When choosing a vector, the method of joining the DNA fragments to the vector and the means of getting the recombinant molecules into the host cell are two main considerations. In practice the host/vector systems in *E. coli* are well defined, so it is a relatively straightforward task to choose the best combination, given the type of fragments to be cloned and the desired outcome of the experiment. However, the great variety of vectors, host cells and cloning kits available from suppliers can be confusing to the first-time gene manipulator, and often a recommendation from an experienced colleague is the best way to proceed.

In devising a cloning strategy all the points mentioned above have to be considered. Often there will be no ideal solution to a particular problem, and a compromise will have to be accepted. By keeping the overall aim of the experiments in mind, the researcher can minimise the effects of such compromises and choose the most efficient cloning route.

6.2 Cloning from mRNA

Each type of cell in a multicellular organism will produce a range of mRNA molecules. In addition to the expression of general 'housekeeping' genes

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

Table 6.1. *m*RNA *abundance classes*

Source	Number of different mRNAs	Abundance (molecules/cells)
Mouse liver cytoplasmic poly(A)$^+$ RNA	9	12 000
	700	300
	11 500	15
Chick oviduct polysomal poly(A)$^+$ RNA	1	100 000
	7	4000
	12 500	5

Note: The diversity of mRNAs is indicated by the number of different mRNA molecules. There is one mRNA that is present in chick oviduct cells at a very high level (100 000 molecules per cell). This mRNA encodes ovalbumin, the major egg white protein. Source: After Old & Primrose (1989), *Principles of Gene Manipulation*, 4th edition, Blackwell. Mouse data from Young *et al.* (1976), *Biochemistry* **15**, 2823–2828, copyright (1976) American Chemical Society. Chick data from Axel *et al.* (1976), Cell **11**, 247–254, copyright (1976) Cell Press. Reproduced with permission.

whose products are required for basic cellular metabolism, cells exhibit tissue-specific gene expression. Thus liver cells, kidney cells, skin cells, etc., will each synthesise a different spectrum of tissue-specific proteins (and hence mRNAs). In addition to the **diversity** of mRNAs produced by each cell type, there may well be different **abundance classes** of particular mRNAs. This has important consequences for cloning from mRNA, as it is easier to isolate a specific cloned sequence if it is present as a high proportion of the starting mRNA population. Some examples of mRNA abundance classes are shown in Table 6.1.

6.2.1 Synthesis of cDNA

It is not possible to clone mRNA directly, so it has to be converted into DNA before being inserted into a suitable vector. This is achieved using the enzyme reverse transcriptase (RTase; see Section 4.2.2) to produce **complementary DNA** (also known as **copy DNA** or **cDNA**). The classic early method of cDNA synthesis utilises the poly(A) tract at the 3′ end of the mRNA to bind an oligo(dT) primer, which provides the 3′-OH group required by RTase (Fig. 6.2). Given the four dNTPs and suitable conditions, RTase will synthesise a copy of the mRNA to produce a cDNA·mRNA hybrid. The mRNA can be

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

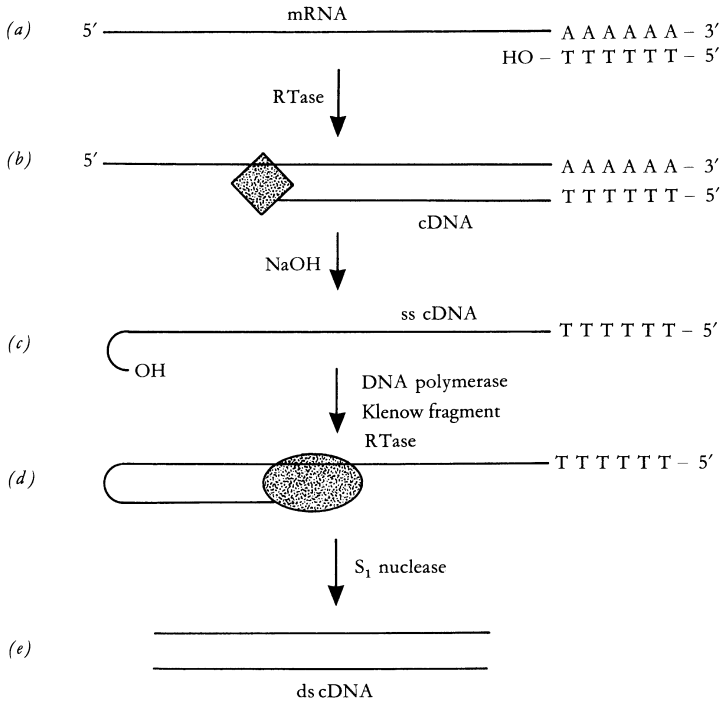


Fig. 6.2. Synthesis of cDNA. Poly(A)$^+$ RNA (mRNA) is used as the starting material. (*a*) A short oligo(dT) primer is annealed to the poly(A) tail on the mRNA, which provides the 3′-OH group for reverse transcriptase to begin copying the mRNA (*b*). The mRNA is removed by alkaline hydrolysis to give a single-stranded cDNA molecule (*c*). This has a short double-stranded hairpin loop structure which provides a 3′-OH terminus for (*d*) second-strand synthesis by a DNA polymerase (T4 DNA polymerase, Klenow fragment, or RTase). (*e*) The double-stranded cDNA is trimmed with S_1 nuclease to produce a blunt-ended ds cDNA molecule. An alternative to the alkaline hydrolysis step is to use RNase H, which creates nicks in the mRNA strand of the mRNA·cDNA hybrid. By using this in conjunction with DNA polymerase I, a nick translation reaction synthesises the second cDNA strand.

removed by alkaline hydrolysis and the single-stranded (ss) cDNA converted into double-stranded (ds) cDNA by using a DNA polymerase. In this second strand synthesis the priming 3′-OH is generated by short hairpin loop regions that form at the end of the ss cDNA. After second strand synthesis, the ds cDNA can be trimmed by S_1 nuclease to give a flush-ended molecule, which can then be cloned in a suitable vector.

Several problems are often encountered in synthesising cDNA using the

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

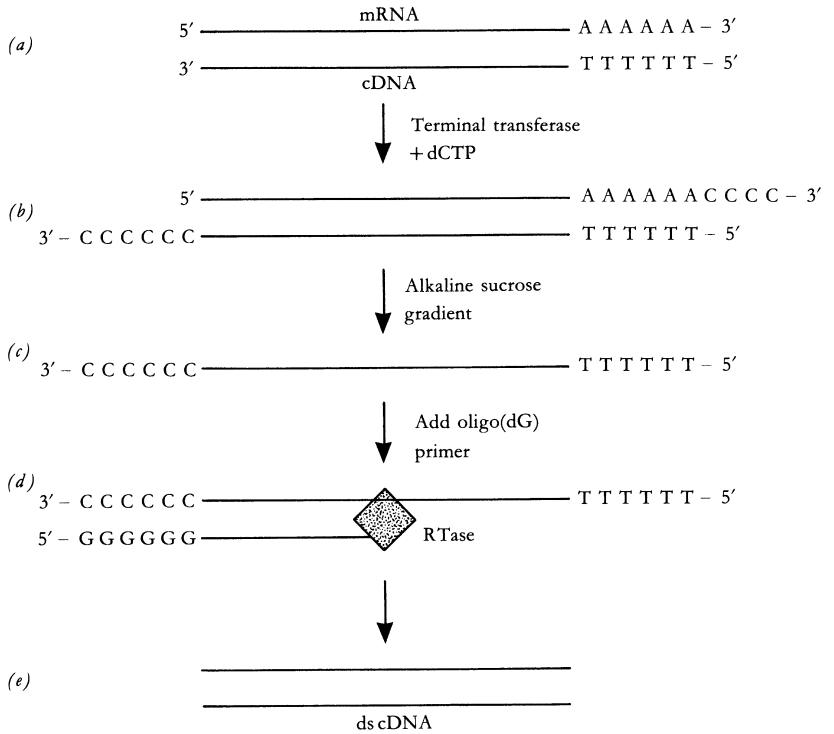


Fig. 6.3. Oligo(dG)-primed second-strand cDNA synthesis. (*a*) First-strand synthesis is as shown in Fig. 6.2, generating an mRNA·cDNA hybrid. (*b*) This is tailed with C residues using terminal transferase. (*c*) Fractionation through an alkaline sucrose gradient hydrolyses the mRNA and permits recovery of full-length cDNA molecules. (*d*) An oligo(dG) primer is annealed to the C tails, and reverse transcriptase used to synthesise the second strand. (*e*) This generates a double-stranded full-length cDNA molecule. From Old and Primrose (1989), *Principles of Gene Manipulation*, Blackwell. (Reproduced with permission.)

method outlined above. Firstly, synthesis of full-length cDNAs may be inefficient, particularly if the mRNA is relatively long. This is a serious problem if expression of the cDNA is required, as it may not contain all the coding sequence of the gene. Such inefficient full-length cDNA synthesis also means that the 3′ regions of the mRNA tend to be over-represented in the cDNA population. Secondly, problems can arise from the use of S_1 nuclease, which may remove some important 5′ sequences when it is used to trim the ds cDNA.

More recent methods for cDNA synthesis overcome the above problems to a great extent, and the original method is now rarely used. One of the sim-

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

plest adaptations involves the use of oligo(dC) tailing to permit oligo(dG)-primed second-strand cDNA synthesis (Fig. 6.3). The dC tails are added to the 3′ termini of the cDNA using the enzyme terminal transferase. This functions most efficiently on accessible 3′ termini, and the tailing reaction therefore favours full-length cDNAs in which the 3′ terminus is not 'hidden' by the mRNA template. The method also obviates the need for S_1 nuclease treatment, and thus full-length cDNA production is enhanced further.

Many suppliers now produce kits for cDNA synthesis. Often these have been optimised for a particular application, and the number of steps involved is usually reduced to a minimum. In many ways the mystique that surrounded cDNA synthesis in the early days has now gone, and the techniques available make full-length cDNA synthesis a relatively straightforward business. The key to success is to obtain good quality mRNA preparations and to take great care in handling these. In particular, contamination with nucleases must be avoided.

Although the poly(A) tract of eukaryotic mRNAs is often used for priming cDNA synthesis, there may be cases where this is not appropriate. Where the mRNA is not polyadenylated, random oligonucleotide primers may be used to initiate cDNA synthesis. Or, if all or part of the amino acid sequence of the desired protein is known, a specific oligonucleotide primer can be synthesised and used to initiate cDNA synthesis. This can be of great benefit in that specific mRNAs may be copied into cDNA, which simplifies the screening procedure when the clones are obtained. An additional possibility with this approach is to use the polymerase chain reaction (PCR; see Chapter 7) to amplify selectively the desired sequence.

Having generated the cDNA fragments, the cloning procedure can begin. Here there is a further choice to be made regarding the vector system – plasmid or phage, or perhaps cosmid or phagemid? Examples of cloning strategies based on the use of plasmid and phage vectors are given below.

6.2.2 Cloning cDNA in plasmid vectors

Although many workers prefer to clone cDNA using a bacteriophage vector system, plasmids are still often used, particularly where isolation of the desired cDNA sequence involves screening a relatively small number of clones. Joining the cDNA fragments to the vector is usually achieved by one of the three methods outlined in Fig. 6.1 for cDNA cloning, these being **blunt-end ligation**, the use of **linker molecules**, and **homopolymer tailing**. Although favoured for cDNA cloning, these methods may also be used with genomic DNA (see Section 6.3). Each of the three methods will be described briefly.

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

Blunt-end ligation is exactly what it says – the joining of DNA molecules with blunt ends, using DNA ligase (see Section 4.3). In cDNA cloning, the blunt ends may arise as a consequence of the use of S_1 nuclease, or they may be generated by filling in the protruding ends with DNA polymerase. The main disadvantage of blunt-end ligation is that it is an inefficient process, as there is no specific intermolecular association to hold the DNA strands together whilst DNA ligase generates the phosphodiester linkages required to produce the recombinant DNA. Thus high concentrations of the participating DNAs must be used, so that the chances of two ends coming together are increased. The effective concentration of DNA molecules in cloning reactions is usually expressed as the concentration of termini, thus one talks about 'picomoles of ends', which can seem rather strange terminology to the uninitiated.

The conditions for ligation of ends must be chosen carefully. In theory, when vector DNA and cDNA are mixed, there are several possible outcomes. The desired result is for one cDNA molecule to join with one vector molecule, thus generating a recombinant with one insert. However, if concentrations are not optimal, the insert or vector DNAs may self-ligate to produce circular molecules, or the insert/vector DNAs may form concatemers instead of bimolecular recombinants. In practice, the vector is often treated with a phosphatase (either BAP or CIP; see Section 4.2.3) to prevent self-ligation, and the concentrations of the vector and insert DNAs are chosen to favour the production of recombinants.

One potential disadvantage of blunt-end ligation is that it may not generate restriction enzyme recognition sequences at the cloning site, thus hampering excision of the insert from the recombinant. This is usually not a major problem, as many vectors now have a series of restriction sites clustered around the cloning site. Thus DNA inserted by blunt-end ligation can often be excised by using one of the restriction sites in the cluster. Another approach involves the use of **linkers**, which are self-complementary oligomers that contain a recognition sequence for a particular restriction enzyme. One such sequence would be 5′-CCGAATTCGG-3′, which in double-stranded form will contain the recognition sequence for *Eco*RI (GAATTC). Linkers are synthesised chemically, and can be added to cDNA by blunt-end ligation (Fig. 6.4). When they have been added, the cDNA/linker is cleaved with the linker-specific restriction enzyme, thus generating sticky ends prior to cloning. This can pose problems if the cDNA contains sites for the restriction enzyme used to cleave the linker, but these may be overcome by using a methylase to protect any internal recognition sites from digestion by the enzyme.

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

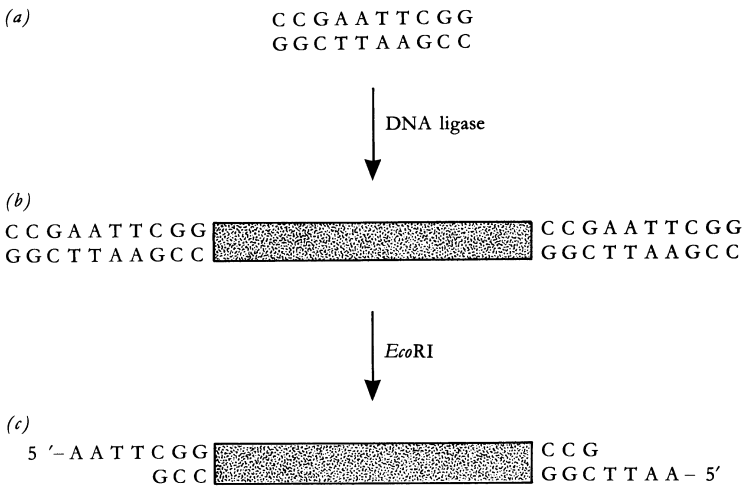


Fig. 6.4. Use of linkers. (*a*) The 10-mer 5′-CCGAATTCGG-3′ contains the recognition site for EcoRI. (*b*) The linker is added to blunt-ended DNA using DNA ligase. (*c*) The construct is then digested with *Eco*RI, which cleaves the linker to generate protruding 5′ termini. Redrawn from Winnacker (1987), *From Genes to Clones*, VCH. (Reproduced with permission.)

A second approach to cloning by addition of sequences to the ends of DNA molecules involves the use of **adaptors** (Fig. 6.5). These are single-stranded non-complementary oligomers that may be used in conjunction with linkers. When annealed together, a linker/adaptor with one blunt end and one sticky end is produced, which can be added to the cDNA to provide sticky-end cloning without digestion of the linkers.

The use of **homopolymer tailing** has proved to be a popular and effective means of cloning cDNA. In this technique, the enzyme terminal transferase (see Section 4.2.3) is used to add homopolymers of dA, dT, dG or dC to a DNA molecule. Early experiments in recombinant production used dA tails on one molecule and dT tails on the other, although the technique is now most often used to clone cDNA into the *Pst*I site of a plasmid vector by dG·dC tailing. Homopolymers have two main advantages over other methods of joining DNAs from different sources. Firstly, they provide longer regions for **annealing** DNAs together than, for example, cohesive termini produced by restriction enzyme digestion. This means that **ligation** need not be carried out *in vitro*, as the cDNA·vector hybrid is stable enough to survive introduction into the host cell, where it is ligated *in vivo*. A second advantage is specificity. As the vector and insert cDNAs have different but complementary 'tails',

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

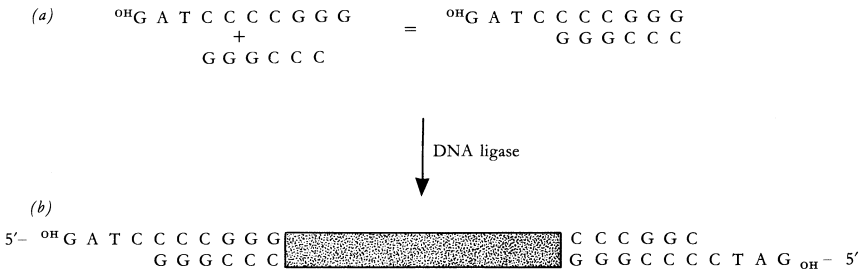


Fig. 6.5. Use of adaptors. In this example a *Bam*HI adaptor (5′-GATCCCCGGG-3′) is annealed with a single-stranded *Hpa*II linker (3′-GGGCCC-5′) to generate a double-stranded sticky-ended molecule, as shown in (*a*). This is added to blunt-ended DNA using DNA ligase. The DNA therefore gains protruding 5′ termini without the need for digestion with a restriction enzyme, as shown in (*b*). The 5′ terminus of the adaptor can be dephosphorylated to prevent self ligation. Redrawn from Winnacker (1987), *From Genes to Clones*, VCH. (Reproduced with permission.)

there is little chance of self-annealing, and the generation of bimolecular recombinants is favoured over a wider range of effective concentrations that is the case for other annealing/ligation reactions.

An example of the use of homopolymer tailing is shown in Fig. 6.6. The vector is cut with *Pst*I and tailed by terminal transferase in the presence of dGTP. This produces dG tails. The insert DNA is tailed with dC in a similar way, and the two can then be annealed. This regenerates the original *Pst*I site, which enables the insert to be cut out of the recombinant using this enzyme.

Introduction of cDNA·plasmid recombinants into suitable *E. coli* hosts is achieved by transformation (Section 5.5.1), and the desired transformants can then be selected by the various methods available (see Chapter 8).

6.2.3 Cloning cDNA in bacteriophage vectors

Although plasmid vectors have been used extensively in cDNA cloning protocols, there are situations where they may not be appropriate. If a large number of recombinants is required, as might be the case if a low-abundance mRNA was to be cloned, phage vectors may be more suitable. The chief advantage here is that packaging *in vitro* may be used to generate the recombinant phage, which greatly increases the efficiency of the cloning process. In addition, it is much easier to store and handle large numbers of phage clones than is the case for bacterial colonies carrying plasmids. Given that isolation of a cDNA

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

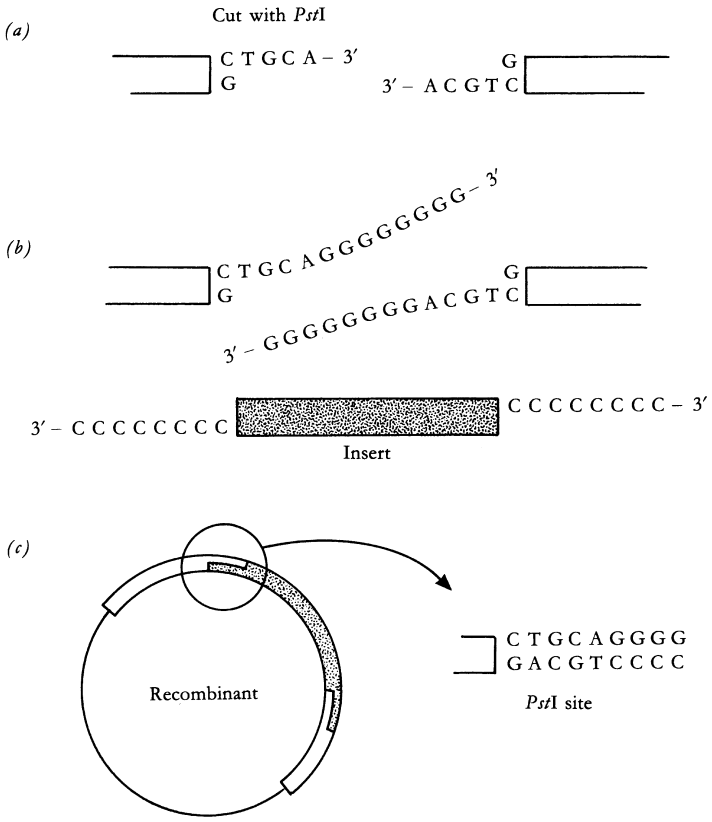


Fig. 6.6. Homopolymer tailing. (*a*) The vector is cut with *PstI*, which generates protruding 3′-OH termini. (*b*) The vector is then tailed with dG residues using terminal transferase. The insert DNA is tailed with dC residues in a similar way. (*c*) The dC and dG tails are complementary and the insert can therefore be annealed with the vector to generate a recombinant. The *Pst*I sites are regenerated at the ends of the insert DNA, as shown.

clone of a rare mRNA species may require screening hundreds of thousands of independent clones, ease of handling becomes a major consideration.

Cloning cDNA in phage λ vectors is, in principle, no different to cloning any other piece of DNA. However, the vector has to be chosen carefully, as cDNA cloning has slightly different requirements than genomic DNA cloning in λ vectors (see Section 6.3). Generally cDNAs will be much shorter than genomic DNA fragments, so an insertion vector is usually chosen. Vectors such as λgt10 and Charon 16A (Section 5.3.2) are suitable, with cloning capacities of some 7.6 and 9.0 kb respectively. The cDNA may be size-fractionated prior to cloning, to remove short cDNAs that may not be representative full-length copies of the

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

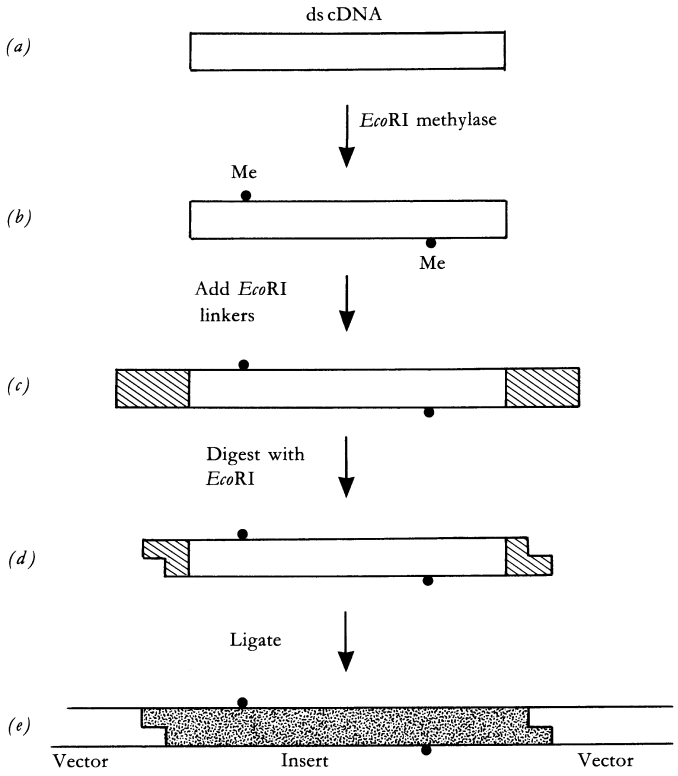


Fig. 6.7. Cloning cDNA in λ vectors using linkers. (*a*) The ds cDNA is treated with *Eco*RI methylase, which (*b*) methylates any internal *Eco*RI recognition sequences. (*c*) *Eco*RI linkers are then added to the ends of the methylated cDNA, and the linkers digested with *Eco*RI. (*d*) The methylation prevents digestion at internal sites, and the result is a cDNA with *Eco*RI cohesive ends. (*e*) This can be ligated into the *Eco*RI site of a λ vector such as λgt10.

mRNA. In the case of vectors such as λgt10, cDNA is usually ligated into the *Eco*RI site using linkers, as shown in Fig. 6.7. The recombinant DNA is packaged *in vitro* and plated on a suitable host for selection and screening.

6.3 Cloning from genomic DNA

Although cDNA cloning is an extremely useful branch of gene manipulation technology, there are certain situations where cDNAs will not provide the answers to the questions that are being posed. If, for example, the overall

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

structure of a particular **gene** is being investigated (as opposed to its RNA transcript), the investigator may wish to determine if there are introns present. He or she will probably also wish to examine the control sequences responsible for regulating gene expression, and these will not be present in the processed mRNA molecule that is represented by a cDNA clone. In such a situation clones generated from genomic DNA must be isolated. This presents a slightly different set of problems than those involved in cloning cDNA, and therefore requires a different cloning strategy.

6.3.1 Genomic libraries

Cloning DNA, by whatever method, gives rise to a population of recombinant DNA molecules, often in plasmid or phage vectors, maintained either in bacterial cells or as phage particles. A collection of independent clones is termed a **clone bank** or **library**. The term **genomic library** is often used to describe a set of clones representing the entire genome of an organism, and the production of such a library is usually the first step in isolating a DNA sequence from an organism's genome.

What are the characteristics of a good genomic library? In theory, a genomic library should represent the entire genome of an organism as a set of overlapping cloned fragments, produced in a random manner, and maintained in a stable form with no misrepresentation of sequences. The systems available for producing genomic libraries essentially fulfil these requirements, although some compromise may be necessary during the cloning process.

The first consideration in constructing a genomic library is the number of clones required. This depends on a variety of factors, the most obvious one being the size of the genome. Thus a small genome such as that of *E. coli* will require many fewer clones than a more complex one such as the human genome. The type of vector to be used also has to be considered, which will determine size of fragments that can be cloned. In practice, library size can be calculated quite simply on the basis of the probability of a particular sequence being represented in the library. There is a formula that takes account of all the factors and produces a 'number of clones' value. The formula is:

$$N = \ln(1 - P)/\ln(1 - a/b)$$

where N is the number of clones required, P is the desired probability of a particular sequence being represented (typically set at 0.95 or 0.99), a is the average size of the DNA fragments to be cloned and b is the size of the genome (expressed in the same units as a).

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

Table 6.2. *Genomic library sizes for various organisms*

Organism	Genome size (kb)	No. clones N, $P = 0.95$	
		20kb inserts	45 kb inserts
Escherichia coli (bacterium)	4.0×10^3	6.0×10^2	2.7×10^2
Saccharomyces cerevisiae (yeast)	1.4×10^4	2.1×10^3	9.3×10^2
Arabidopsis thaliana (simple higher plant)	7.0×10^4	1.1×10^4	4.7×10^3
Drosophila melanogaster (fruit fly)	1.7×10^5	2.5×10^4	1.1×10^4
Stronglyocentrotus purpuratus (sea urchin)	8.6×10^5	1.3×10^5	5.7×10^4
Homo sapiens (human)	3.0×10^6	4.5×10^5	2.0×10^5
Triticum aestivum (hexaploid wheat)	1.7×10^7	2.5×10^6	1.1×10^6

Note: The number of clones (N) required for a probability (P) of 95% that a given sequence is represented is shown for various organisms. The genome sizes of the organisms are given (haploid genome size, if appropriate). Two values of N are shown, for 20kb inserts (λ replacement vector size) and 45 kb inserts (cosmid vectors). The values should be considered as minimum estimates, as strictly speaking the calculation assumes: (i) that the genome size is known accurately, (ii) that the DNA is fragmented in a totally random manner for cloning, (iii) that each recombinant DNA molecule will give rise to a single clone, (iv) that the efficiency of cloning is the same for all fragments, and (v) that diploid organisms are homozygous for all loci. These assumptions are usually not all valid for a given experiment.

By using this formula, it is possible to determine the magnitude of the task ahead, and to plan a cloning strategy accordingly. Some genome sizes and their associated library sizes are shown in Table 6.2. These library sizes should be considered as minimum values, as the generation of cloned fragments may not provide a completely random and representative set of clones in the library. Thus, for a human genomic library, we are talking of some 10^6 clones or more in order to be reasonably sure of isolating a particular single-copy gene sequence.

When dealing with this size of library, phage or cosmid vectors are usually

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

essential, as the cloning capacity and efficiency of these vectors is much greater than that of plasmid vectors. Although cosmids, with the potential to clone fragments of up to 47 kb, would seem to be the better choice, λ replacement vectors are often used for library construction. This is because they are easier to use than cosmid vectors, and this outweighs the disadvantage of having only half the cloning capacity. In addition, the techniques for screening phage libraries are now routine and have been well characterised. This is an important consideration, particularly where workers new to the technology wish to use gene manipulation in their research. Alternatively, artificial chromosome vectors such as BACs or YACs may be used to clone large DNA fragments.

6.3.2 Preparation of DNA fragments for cloning

One of the most important aspects of library production is the generation of genomic DNA fragments for cloning. If a λ replacement vector such as EMBL4 is to be used, the maximum cloning capacity will be around 23 kb. Thus fragments of this size must be available for the production of recombinants. In practice a range of fragment sizes is used, often between 17 and 23 kb for a vector such as EMBL4. It is important that smaller fragments are not ligated into the vector, as there is the possibility of multiple inserts which could arise by ligation of small non-contiguous DNA sequences into the vector.

There are two main considerations when preparing DNA fragments for cloning, these being: (i) the molecular weight of the DNA after isolation from the organism, and (ii) the method used to fragment the DNA. For a completely random library, the starting material should be very high molecular weight DNA, and this should be fragmented by a totally random (i.e. sequence-independent) method. Isolation of DNA in excess of 100 kb in length is desirable, and this in itself can pose technical difficulties where the type of source tissue does not permit gentle disruption of cells. In addition, pipetting and mixing solutions of high molecular weight DNA can cause shearing of the molecules, and great care must be taken when handling the preparations.

Assuming that sufficient DNA of 100 kb is available, fragmentation can be carried out. This is usually followed by a size-selection procedure to isolate fragments in the desired range of sizes. Fragmentation can be achieved either by mechanical shearing or by partial digestion with a restriction enzyme. Although mechanical shearing (by forcing the DNA through a syringe needle,

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

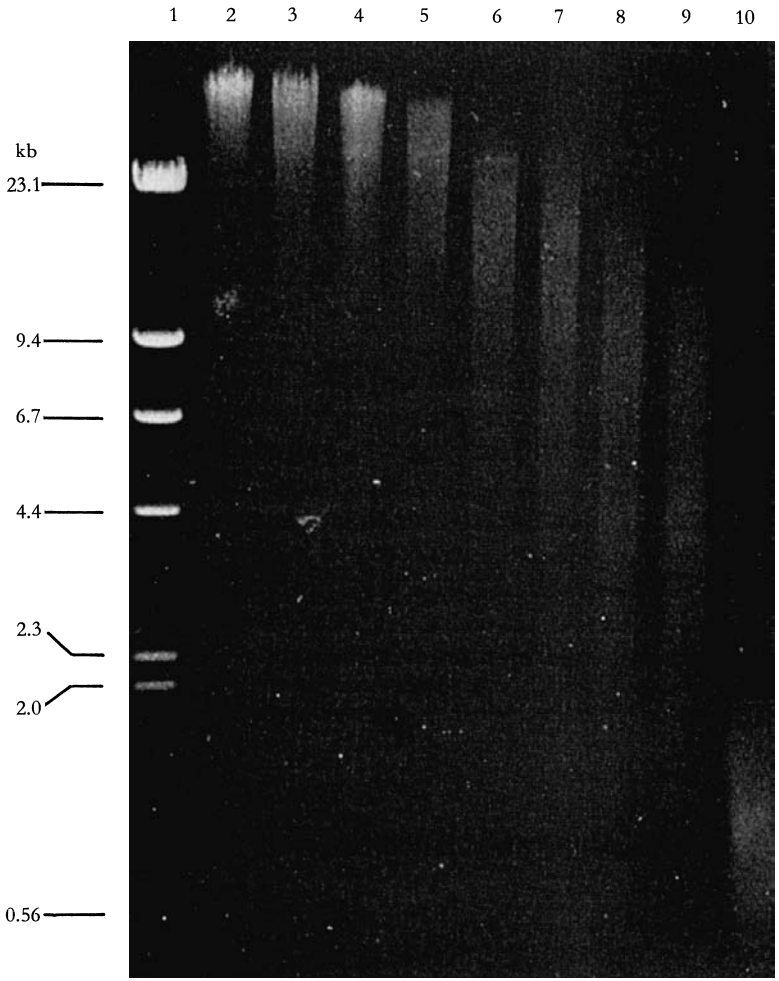


Fig. 6.8. Partial digestion and fractionation of genomic DNA. High molecular weight genomic DNA was digested with various concentrations of the restriction enzyme *Sau3*A. Samples from each digest were run on a 0.7% (w/v) agarose gel and stained with ethidium bromide. Lane 1 shows λ *Hind*III markers, sizes as indicated. Lanes 2 to 10 show the effects of increasing concentrations of restriction enzyme in the digestions. As the concentration of enzyme is increased, the DNA fragments generated are smaller. From this information the optimum concentration of enzyme to produce fragments of a certain size distribution can be determined. These can then be run on a gel (as here) and isolated prior to cloning. (Photograph courtesy of Dr N. Urwin.)

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

or by sonication) will generate random fragments, it will not produce DNA with cohesive termini. Thus further manipulation such as trimming or filling in the ragged ends of the molecules will be required before the DNA can be joined to the vector, usually with linkers, adaptors or homopolymer tails (see Fig. 6.1). In practice these additional steps are often considered undesirable, and fragmentation by partial restriction digestion is used extensively in library construction. However, this is not a totally sequence-independent process, as the occurrence of restriction enzyme recognition sites is clearly sequence-dependent. Partial digestion is therefore something of a compromise, but careful design and implementation of the procedure can overcome most of the disadvantages.

If a restriction enzyme is used to digest DNA to completion, the fragment pattern will obviously depend on the precise location of recognition sequences. This approach therefore has two drawbacks. Firstly, a six-cutter such as *Eco*RI will have recognition sites on average about once every 4096 bp, which would produce fragments that are too short for λ replacement vectors. Secondly, any sequence bias, perhaps in the form of repetitive sequences, may skew the distribution of recognition sites for a particular enzyme. Thus some areas of the genome may contain few sites, whilst others have an over-abundance. This means that a complete digest will not be suitable for generating a representative library. If, however, a **partial** digest is carried out using an enzyme that cuts frequently (e.g. a four-cutter such as *Sau*3A, which cuts on average once every 256 bp), the effect is to produce a collection of fragments that are essentially random. This can be achieved by varying the enzyme concentration or the time of digestion, and a test run will produce a set of digests which contain different fragment size distribution profiles, as shown in Fig. 6.8.

6.3.3 Ligation, packaging and amplification of libraries

Having established the optimum conditions for partial digestion, a sample of DNA can be prepared for cloning. After digestion the sample is fractionated, either by density gradient centrifugation or by electrophoresis. Fragments in the range 17–23 kb can then be selected for ligation. If *Sau*3A (or *Mbo*I, which has the same recognition sequence) has been used as the digesting enzyme, the fragments can be inserted into the *Bam*HI site of a vector such as EMBL4, as the ends generated by these enzymes are complementary (Fig. 6.9). The insert DNA can be treated with phosphatase to reduce self-ligation or concatemer formation, and the vector can be digested with *Bam*HI and *Sal*I to

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

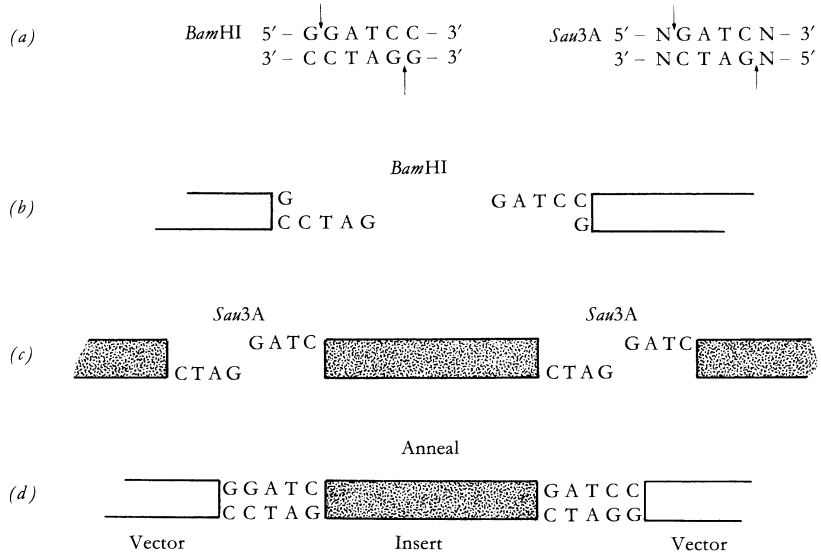


Fig. 6.9. Cloning *Sau*3A fragments into a *Bam*HI site. (*a*) The recognition sequences and cutting sites for *Bam*HI and *Sau*3A. In the *Sau*3A site, N is any base. (*b*) Vector DNA cut with *Bam*HI generates 5′ protruding termini with the four-base sequence 5′-GATC-3′. (*c*) Insert DNA cut with *Sau*3A also generates identical four-base overhangs. (*d*) Thus DNA cut with *Sau*3A can be annealed to *Bam*HI cohesive ends to generate a recombinant DNA molecule.

generate the cohesive ends for cloning and to isolate the stuffer fragment and prevent it from re-annealing during ligation. The *Eco*RI site in the vector can be used to excise the insert after cloning. Ligation of DNA into EMBL4 is summarised in Fig. 6.10.

When ligation is carried out, concatemeric recombinant DNA molecules are produced, which are suitable substrates for packaging *in vitro*, as shown in Fig. 6.11. This produces what is known as a **primary library**, which consists of individual recombinant phage particles. Whilst this is theoretically the most useful type of library in terms of isolation of a specific sequence, it is a finite resource. Thus a primary library is produced, screened and then discarded. If the sequence of interest has not been isolated, more recombinant DNA will have to be produced and packaged. Whilst this may not be a problem, there are occasions where a library may be screened for several different genes, or may be sent to different laboratories. In these cases it is therefore necessary to **amplify** the library. This is achieved by plating the packaged phage on a suitable host strain of *E. coli*, and then resuspending the plaques by gently washing the plates with a buffer solution.

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

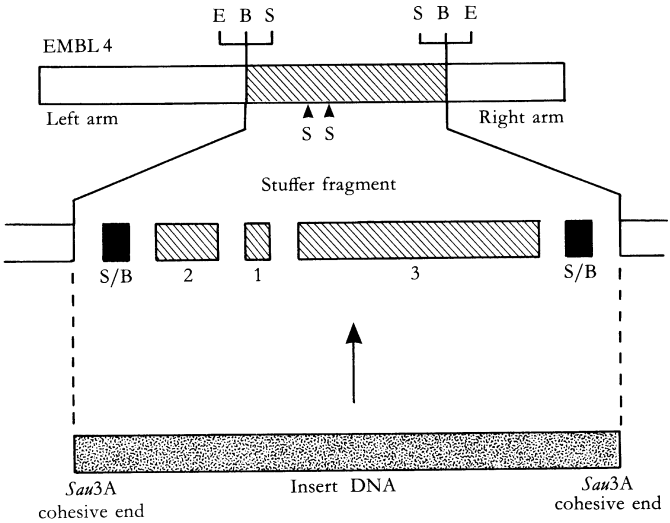


Fig. 6.10. Ligation of *Sau*3A-cut DNA into the λ replacement vector EMBL4. Sites on the vector are *Eco*RI (E), *Bam*HI (B) and *Sal*I (S). The vector is cut with *Bam*HI and *Sal*I, which generates five fragments from the stuffer fragment (hatched in top panel). Removal of the very short *Sal*I/*Bam*HI fragments (filled boxes) prevents the stuffer fragment from re-annealing. In addition, the two internal *Sal*I sites cleave the stuffer fragment, producing three *Sal*I/*Sal*I fragments (1 to 3). If desired, the short fragments can be removed from the preparation by precipitation with isopropanol, which leaves the small fragments in the supernatant. On removal of the stuffer, *Sau*3A-digested insert DNA can be ligated into the *Bam*HI site of the vector (see Fig. 6.9).

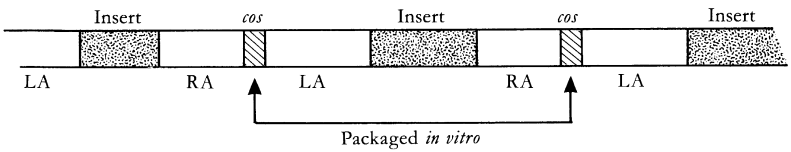


Fig. 6.11. Concatemeric recombinant DNA. On ligation of inserts into a vector such as EMBL4, a concatemer is formed. This consists of the left arm of the vector (LA), the insert DNA, and the right arm (RA). These components of the unit are repeated many times and are linked together at the *cos* sites by the cohesive ends on the vector arms. On packaging *in vitro*, the recombinant genomes are cut at the *cos* sites and packaged into phage heads.

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

The resulting phage suspension can be stored almost indefinitely, and will provide enough material for many screening and isolation procedures.

Although amplification is a useful step in producing stable libraries, it can lead to skewing of the library. Some recombinant phage may be lost, perhaps due to the presence of repetitive sequences in the insert, which can give rise to recombinational instability. This can be minimised by plating on a recombination-deficient host strain. Some phage may exhibit differential growth characteristics, which may cause particular phage to be either over- or under-represented in the amplified library, and this may mean that a greater number of plaques have to be screened in order to isolate the desired sequence.

6.4 Advanced cloning strategies

In Sections 6.2 and 6.3 I examined cDNA and genomic DNA cloning strategies, using basic plasmid and phage vectors in *E. coli* hosts. These approaches have proved to be both reliable and widely applicable, and still represent a major part of the technology of gene manipulation. However, advances made over the past few years have increased the scope (and often the complexity!) of cloning procedures. Such advances include more sophisticated vectors for *E. coli* and other hosts, increased use of expression vectors, and novel approaches to various technical problems, including the extensive use of PCR technology. Some examples of more advanced cloning strategies are discussed below.

6.4.1 Synthesis and cloning of cDNA

An elegant scheme for generating cDNA clones was developed by Hiroto Okayama and Paul Berg in 1982. In their method the plasmid vector itself is used as the **priming** molecule, and the mRNA is annealed to this for cDNA synthesis. A second **adaptor** molecule is required to complete the process. Both adaptor and primer are based on pBR322, with additional sequences from the SV40 virus. Preparation of the vector and adaptor molecules involves restriction digestion, tailing with oligo(dT) and dG, and purification of the fragments to give the molecules shown in Fig. 6.12. The mRNA is then annealed to the plasmid and the first cDNA strand synthesised and tailed with dC. The terminal vector fragment (which is also tailed during this procedure) is removed and the adaptor added to circularise the vector prior to synthesis

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

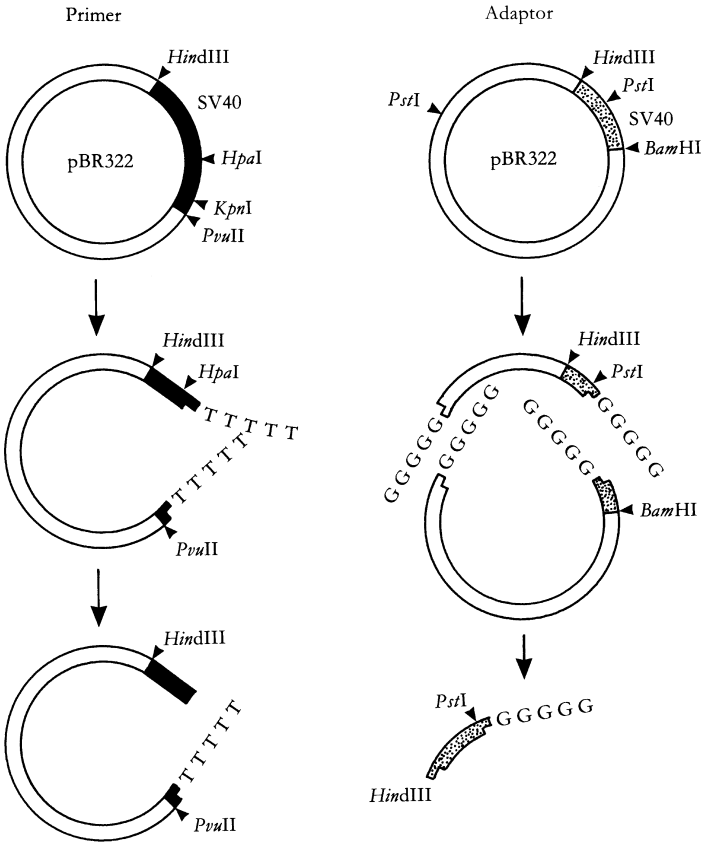


Fig. 6.12. Preparation of vector and adaptor molecules for Okayama and Berg cDNA cloning. The vector is made up from pBR322 plus parts of the SV40 genome (solid or shaded in the diagram). For the primer, the vector is cut with *Kpn*I and tailed with dT residues. It is then digested with *Hpa*I to create a vector in which one end is tailed. The adaptor molecule is generated by cutting the adaptor plasmid with *Pst*I, which generates two fragments. These are tailed with dG residues and digested with *Hind*III to produce the adaptor molecule itself, which therefore has a *Hind*III cohesive end in addition to the dG tail. The fragment is purified for use in the protocol (see Fig. 6.13). From Old and Primrose (1989), *Principles of Gene Manipulation*, Blackwell. (Reproduced with permission.)

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

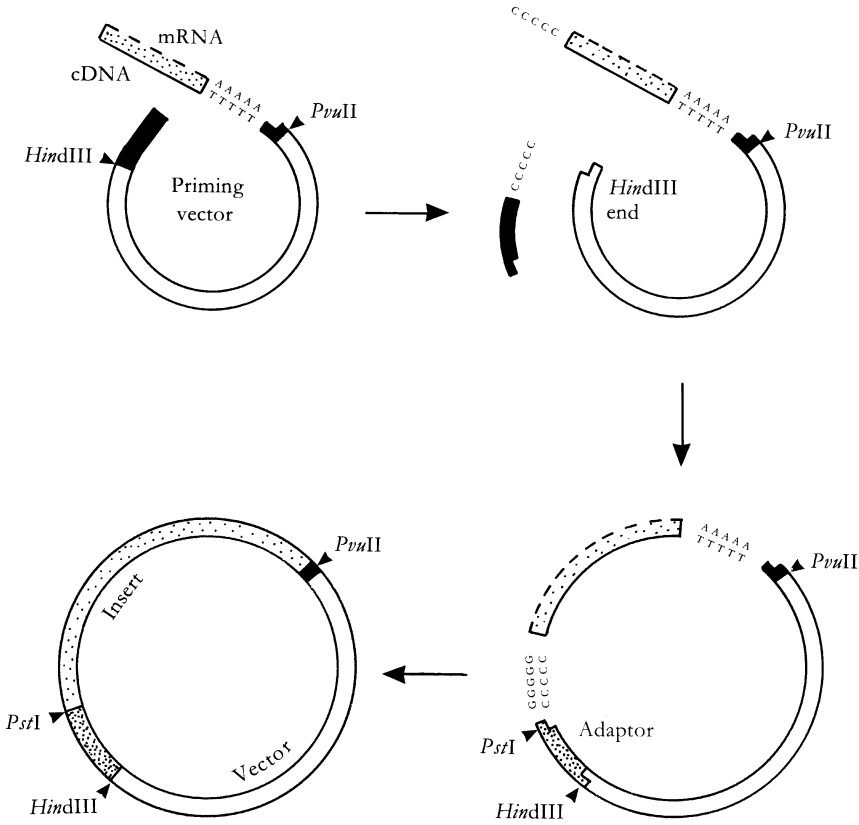


Fig. 6.13. Okayama and Berg cDNA cloning protocol. The mRNA is annealed to the dT-tailed priming vector *via* the poly(A) tail, and the first cDNA strand synthesised using reverse transcriptase. The cDNA is then tailed with dC residues and the dC-tailed vector fragment removed by digestion with *Hind*III. The cDNA is annealed to the dG tailed adaptor molecule, which is ligated into the vector using the cohesive *Hind*III ends on the vector and adaptor molecules. Finally the mRNA is displaced and the second cDNA strand synthesised using RNase H and DNA polymerase I to generate the complete vector/insert recombinant. From Old and Primrose (1989), *Principles of Gene Manipulation*, Blackwell. (Reproduced with permission.)

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

of the second strand of the cDNA. Second-strand synthesis involves the use of RNase H, DNA polymerase I and DNA ligase in a strand-replacement reaction which converts the mRNA·cDNA hybrid into ds cDNA and completes the ligation of the ds cDNA into the vector. The end result is that recombinants are generated in which there is a high proportion of full-length cDNAs. The Okayama and Berg method is summarised in Fig. 6.13.

6.4.2 Expression of cloned DNA molecules

Many of the routine manipulations in gene cloning experiments do not require expression of the cloned DNA. However, there are certain situations in which some degree of genetic expression is needed. A transcript of the cloned sequence may be required for use as a probe, or a protein product (requiring transcription and translation) may be required as part of the screening process used to identify the cloned gene. Another common biotechnological application is where the recombinant DNA is used to produce a protein of commercial value. If eukaryotic DNA sequences are cloned, post-transcriptional and post-translational modifications may be required, and the type of host/vector system that is used is therefore very important in determining whether or not such sequences will be expressed effectively. The problem of RNA processing in prokaryotic host organisms may be obviated by cloning cDNA sequences, and this is the most common approach where expression of eukaryotic sequences is desired. In this section I consider some aspects of cloning cDNAs for expression, concentrating mainly on the characteristics of the vector/insert combination that enable expression to be achieved. Further discussion of the topic is presented in Chapters 9 and 10.

Assuming that a functional cDNA sequence is available, a suitable host/vector combination must be chosen. The host cell type will usually have been selected by considering aspects such as ease of use, fermentation characteristics or the ability to secrete proteins derived from cloned DNA. However, for a given host cell, there may be several types of expression vector, including both plasmid and (for bacteria) phage-based examples. In addition to the normal requirements such as restriction site availability and genetic selection mechanisms, a key feature of expression vectors is the type of **promoter** that is used to direct expression of the cloned sequence. Often the aim will be to maximise the expression of the cloned sequence, so a vector with a highly efficient promoter is chosen. Such promoters are often termed **strong** promoters. However, if the product of the cloned gene is toxic to the cell, a

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

Table 6.3. *Promoters used in expression vectors*

Organism	Gene promoter	Induction by
E. coli	*lac* operon	IPTG
	trp operon	β-Indolylacetic acid
	λP_L	Temperature-sensitive λ cl protein
A. nidulans	Glucoamylase	Starch
S. cerevisiae	Acid phosphatase	Phosphate depletion
	Alcohol dehydrogenase	Glucose depletion
	Galactose utilization	Galactose
	Metallothionein	Heavy metals
T. reesei	Cellobiohydrolase	Cellulose
Mouse	Metallothionein	Heavy metals
Human	Heat-shock protein	Temperature >40°C

Note: Some examples of various promoters that can be used in expression vectors are given, with the organism from which the gene promoter is taken. The conditions under which gene expression is induced from such promoters are also given.
Source: Collated from Brown (1990). *Gene Cloning*, Chapman & Hall; and Old & Primrose (1989). *Principles of Gene Manipulation*, Blackwell. Reproduced with permission.

weak promoter may be required to avoid cell death due to over-expression of the toxic product.

Promoters are regions with a specific base sequence, to which RNA polymerase will bind. By examining the base sequence lying on the 5′ (upstream) side of the coding regions of many different genes, the types of sequences that are important have been identified. Although there are variations, these sequences all have some similarities. The 'best fit' sequence for a region such as a promoter is known as the **consensus sequence**. In prokaryotes there are two main regions that are important. Some 10 base-pairs upstream from the transcription start site (the −10 region, as the T_C start site is numbered +1) there is a region known as the **Pribnow box**, which has the consensus sequence 5′-TATAAT-3′. A second important region is located around position −35, and has the consensus sequence 5′-TTGACA-3′. These two regions form the basis of promoter structure in prokaryotic cells, with the precise sequences found in each region determining the strength of the promoter.

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

Sequences important for transcription initiation in eukaryotes have been identified in much the same way as for prokaryotes. Eukaryotic promoter structure is generally more complex than that found in prokaryotes, and control of initiation of transcription can involve sequences (e.g. enhancers) that may be several hundreds or thousands of base-pairs upstream from the T_C start site. However, there are important motifs closer to the start site. These are a region centred around position −25 with the consensus sequence 5′-TATAAAT-3′ (the **TATA** or **Hogness** box) and a sequence in the −75 region with the consensus 5′-GG(T/C)CAATCT-3′, known as the **CAAT** box.

In addition to the strength of the promoter, it may be desirable to regulate the expression of the cloned cDNA by using promoters from genes that are either **inducible** or **repressible**. Thus some degree of control can be exerted over the transcriptional activity of the promoter; when the cDNA product is required, transcription can be 'switched on' by manipulating the system using an appropriate metabolite. Some examples of promoters used in the construction of expression vectors are given in Table 6.3.

In theory, constructing an expression vector is straightforward once a suitable promoter has been identified. In practice, as is often the case, the process is often highly complex, requiring many manipulations before a functional vector is obtained. The basic vector must carry an origin of replication that is functional in the target host cell, and there may be antibiotic resistance genes or other genetic selection mechanisms present. However, as far as expression of cloned sequences is concerned, it is the arrangement of restriction sites immediately downstream from the promoter that is critical. There must be a unique restriction site for cloning into, and this has to be located in a position where the inserted cDNA sequence can be expressed effectively. This aspect of vector structure is discussed further when the applications of recombinant DNA technology are considered in Part III.

6.4.3 Cloning large DNA fragments in BAC and YAC vectors

Bacterial artificial chromosomes (BACs) and yeast artificial chromosomes (YACs; see Fig. 5.11) can be used to clone very long pieces of DNA. The use of a BAC or YAC vector can reduce dramatically the number of clones needed to produce a representative genomic library for a particular organism, and this is a desirable outcome in itself. A consequence of cloning large pieces of DNA is that physical mapping of genomes is made simpler, as

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

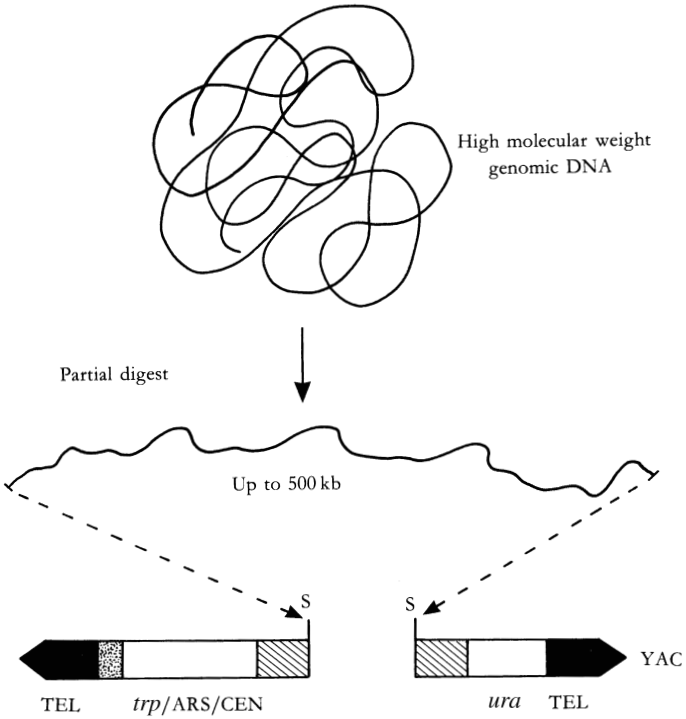


Fig. 6.14. Cloning in a YAC vector. Very large DNA fragments (up to 500 kb) are generated from high molecular weight DNA. The fragments are then ligated into a YAC vector (see Fig. 5.11) that has been cut with *Bam*HI and *Sma*I (S). The construct contains the cloned DNA and the essential requirements for a yeast chromosome, i.e. telomeres (TEL), an autonomous replication sequence (ARS) and a centromere region (CEN). The *trp* and *ura* genes can be used as dual selectable markers to ensure that only complete artificial chromosomes are maintained. From Kingsman and Kingsman (1988), *Genetic Engineering*, Blackwell. (Reproduced with permission.)

there are not as many non-contiguous sequences to fit together in the correct order.

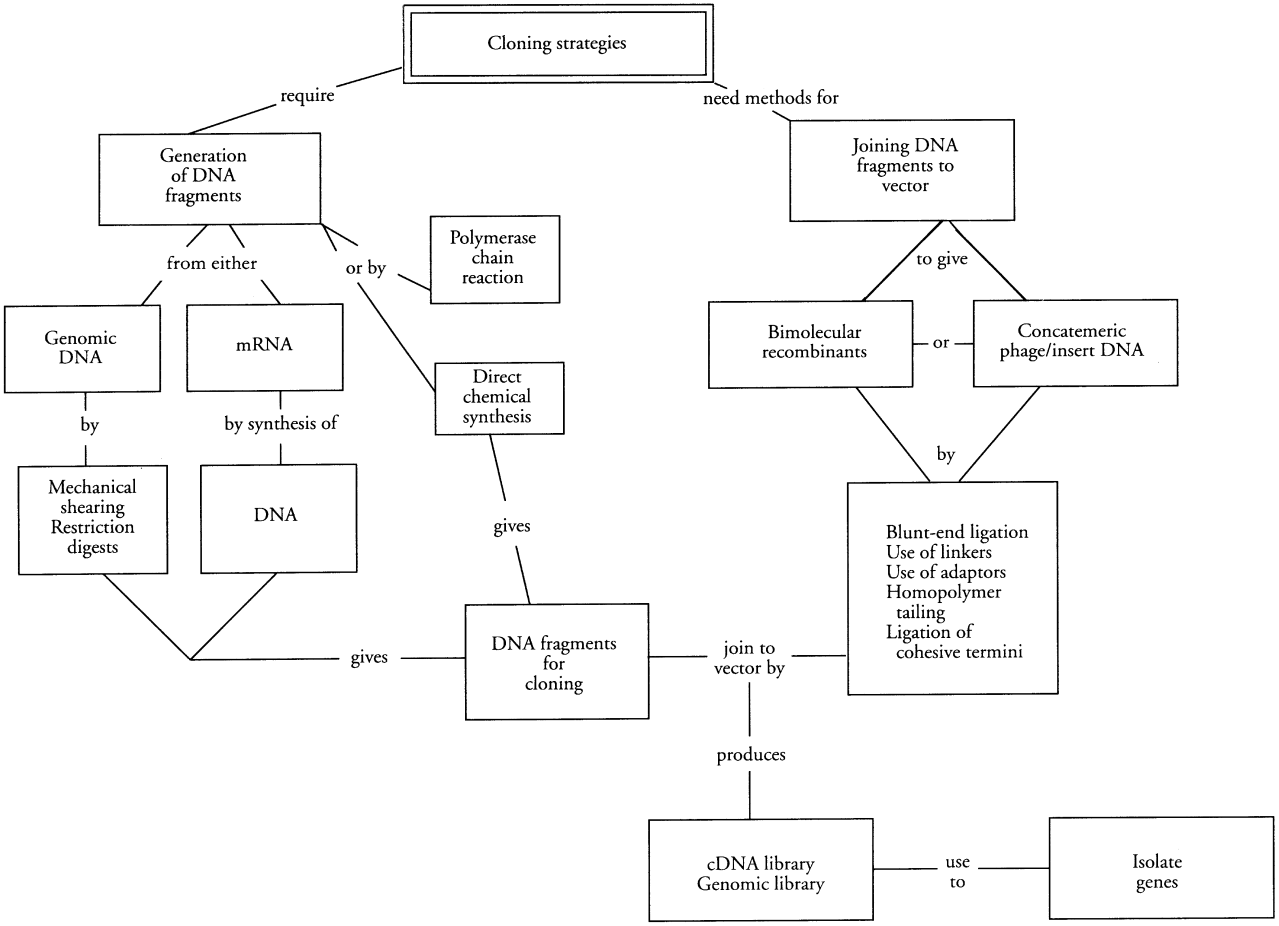
A further advantage of cloning long stretches of DNA stems from the fact that many eukaryotic genes are much larger than the 47 kb or so that can be cloned using cosmid vectors in *E. coli*. Thus with plasmid, phage and cosmid vectors it may be impossible to isolate the entire gene. This makes it difficult to determine gene structure without using several different clones, which is not the ideal way to proceed. The use of BAC or YAC vectors can alleviate

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007

this problem and can enable the structure of large genes to be determined by providing a single DNA fragment to work from.

Let's consider using a YAC vector to clone DNA fragments. In practice, cloning in YAC vectors is similar to other protocols (Fig. 6.14). The vector is prepared by a double restriction digest, which releases the vector sequence between the telomeres and cleaves the vector at the cloning site. Thus two arms are produced, as is the case with phage vectors. Insert DNA is prepared as very long fragments (a partial digest with a six-cutter may be used) and ligated into the cloning site to produce artificial chromosomes. Selectable markers on each of the two arms ensure that only correctly constructed chromosomes will be selected and propagated.

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/CBO9781139168205.007



Concept map 6

Downloaded from https://www.cambridge.org/core. The Librarian-Seeley Historical Library, on 14 Nov 2019 at 07:15:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139168205.007