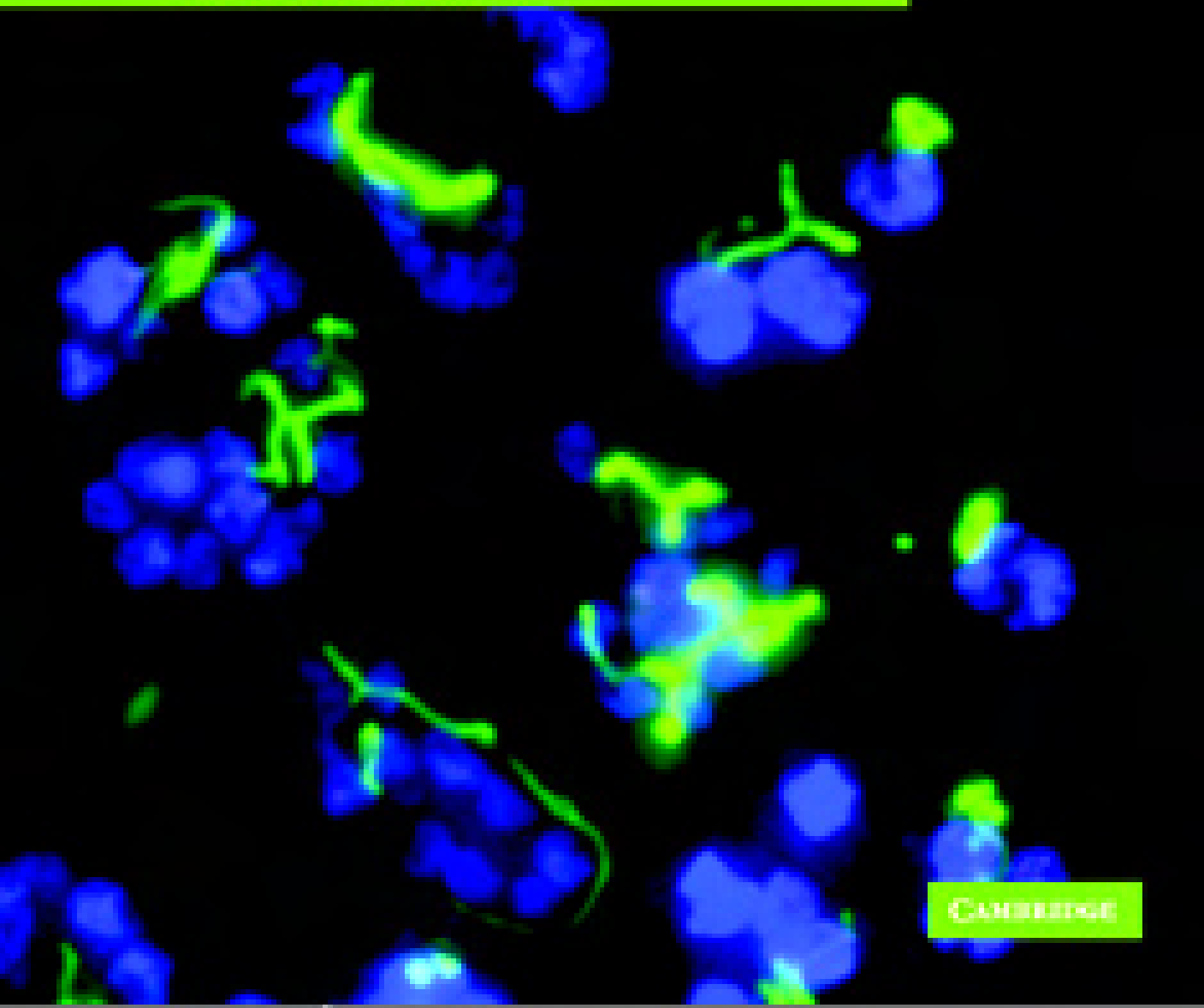
Second Edition

Gene Cloning and Manipulation

Christopher Howe



CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521817936

Chapter 3

Simple cloning

3.1 The basic experiment

3.1.1 Cloning using a simple plasmid

Now that we have surveyed the enzymes and techniques that are used in cloning, we can take an easy example and see how they are applied. One of the simplest exercises is inserting DNA into *E. coli* in a plasmid such as pUC18 (Figure 3.1). Bacterial plasmids are circular dsDNA molecules capable of replicating in bacterial cells. Insertion of DNA into the plasmid allows the DNA to be propagated in a host cell. Molecules that are used for propagating DNA in this way are called vectors. The pUC18 plasmid is one of a series of similar vectors and carries two genes of particular importance. One confers resistance to the antibiotic ampicillin (Figure 3.2), which blocks the cross-linking of the polysaccharide chains in the bacterial cell walls. The lack of cross-linking weakens the bacterial cell wall and the cells eventually lyse. The resistance gene (*bla* or amp^R) encodes a beta-lactamase enzyme that hydrolyses the beta-lactam ring of the antibiotic, inactivating it. The other gene (*lacZ′*) encodes part of

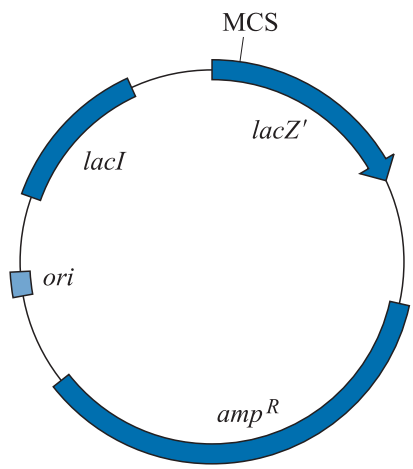


Fig 3.1 Generalized pUC plasmid. It contains an ampicillin resistance gene (amp^R), part of a beta-galactosidase gene (*lacZ′*), the Lac repressor gene (*lacI*) and an orgin of replication (*ori*). The multiple cloning site (MCS) is located in the beta-galactosidase gene.

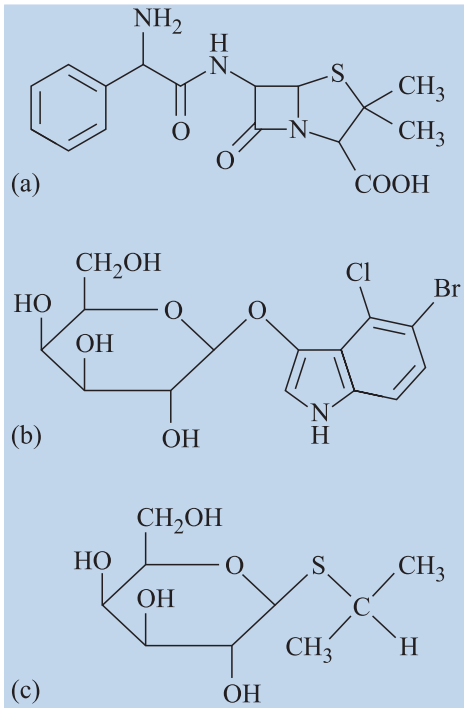


Fig 3.2 Structure of ampicillin (a), X-gal or 5-bromo-4-chloro-3-indolyl-β-D-galactoside (b), and IPTG or isopropyl-β-D-thiogalactoside (c).

a beta-galactosidase enzyme, which normally cleaves disaccharides, such as lactose, into monosaccharides. It can also cleave an artificial substrate 5-bromo-4-chloro-3-indolyl-β-D-galactoside (also known as X-gal), shown in Figure 3.2, to liberate a blue dye. For this reason, the substrate is said to be chromogenic. We will see later why the gene for the whole beta-galactosidase enzyme is not used, but for the moment we can treat the plasmid as though it contained the entire gene. As well as the amp^R and *lacZ′* genes, the plasmid has some less immediately obvious features that we will look at in more detail later, such as an origin of replication. It also has a number of recognition sites for restriction endonucleases, and many of these are located in a region of the *lacZ′* gene known as the multiple cloning site or polylinker.

Suppose we wanted to clone genomic DNA from the organism we were studying into the *Bam*HI site of pUC18, located in the multiple cloning site in the *lacZ′* gene. (We do not have to use the *Bam*HI site. We could use any of the other restriction sites in the multiple cloning site.) The procedure would be as follows. The first stage is to purify genomic DNA from the organism of interest and cut it with the enzyme *Bam*HI. We would also purify some pUC18 and cut it with the same enzyme. Then the *Bam*HI-digested plasmid and the *Bam*HI-digested genomic DNA are mixed and DNA ligase in a suitable buffer is added. *Bam*HI ends will anneal to *Bam*HI ends and the ligase will seal the nicks. That will generate a large number of types of molecule, which may be linear or (topologically) circular.

Linear molecules will not be stably propagated in later stages of the experiment, so we will not consider them further. Different circular molecules that may be formed are shown in Figure 3.3. Intramolecular ligation of plasmid DNA will regenerate the original circular pUC18 plasmid. Intramolecular ligation of genomic DNA will produce circular molecules of genomic DNA, with no plasmid DNA sequences present. A wide range of intermolecular ligations can take place. Two plasmid DNA molecules can be ligated, in a head-to-tail or head-to-head configuration. Two genomic DNA molecules could be ligated, again in either orientation. Most important, a genomic DNA molecule can be ligated to a plasmid DNA molecule. (Of course, a huge range of ligations involving three or more molecules can also take place, but they will be rarer than unimolecular or bimolecular ligations. How rare depends on the concentrations of the species involved.)

The molecules that contain new combinations of sequences, not present before, are termed recombinants. Note that the insertion of DNA into the *lacZ′* gene will destroy the function of that gene, as it disrupts the coding region for the LacZ′ protein; therefore, it is an easy matter to work out what properties each of the plasmid types shown in Figure 3.3 will confer. It is possible that insertion of a short piece of DNA into the *lacZ′* gene might only partially inactivate the gene function (or not affect it at all) if the reading frame were not terminated or shifted. This is relatively uncommon, but the possibility should always be kept in mind.

The next stage is to put the ligation products back into *E. coli*. For an experiment such as this, it would usually be done by chemically induced transformation, or electroporation (Sections 1.5.2 and 1.5.3). An important point is that the uptake of DNA by *E. coli* is rather inefficient, in that the probability of any individual bacterial cell taking up a piece of DNA is very low. For simple transformation procedures, of 10^9 cells treated with 1 μg of supercoiled plasmid DNA, only 10^5 or so will successfully take up a molecule. We say there is a transformation frequency of 10^5 colonies/microgram of DNA. This is low, given that 1 μg of pUC18 contains of the order of 10^{11} molecules. The cells are now plated on agar containing, in this example, ampicillin, together with an inducer of the *lacZ′* gene (isopropyl-thiogalactoside (IPTG), Figure 3.2) and the chromogenic substrate X-gal. The cells that have not taken up a plasmid will be sensitive to the ampicillin and die after a few generations. Those that took up genomic DNA only will also die. Those that took up pUC18 DNA will have acquired an ampicillin resistance gene and be able to grow on the selective medium. They will, therefore, form isolated colonies on the agar if plated at a suitable dilution. Thus, all the members of a colony are derived from one cell and should have acquired a single plasmid. (Because the probability of a cell taking up one DNA molecule is low, the probability of a cell taking up two might be expected to be very low indeed. Actually, that is likely not to be the case. Although only a tiny fraction of cells is able to take up DNA,

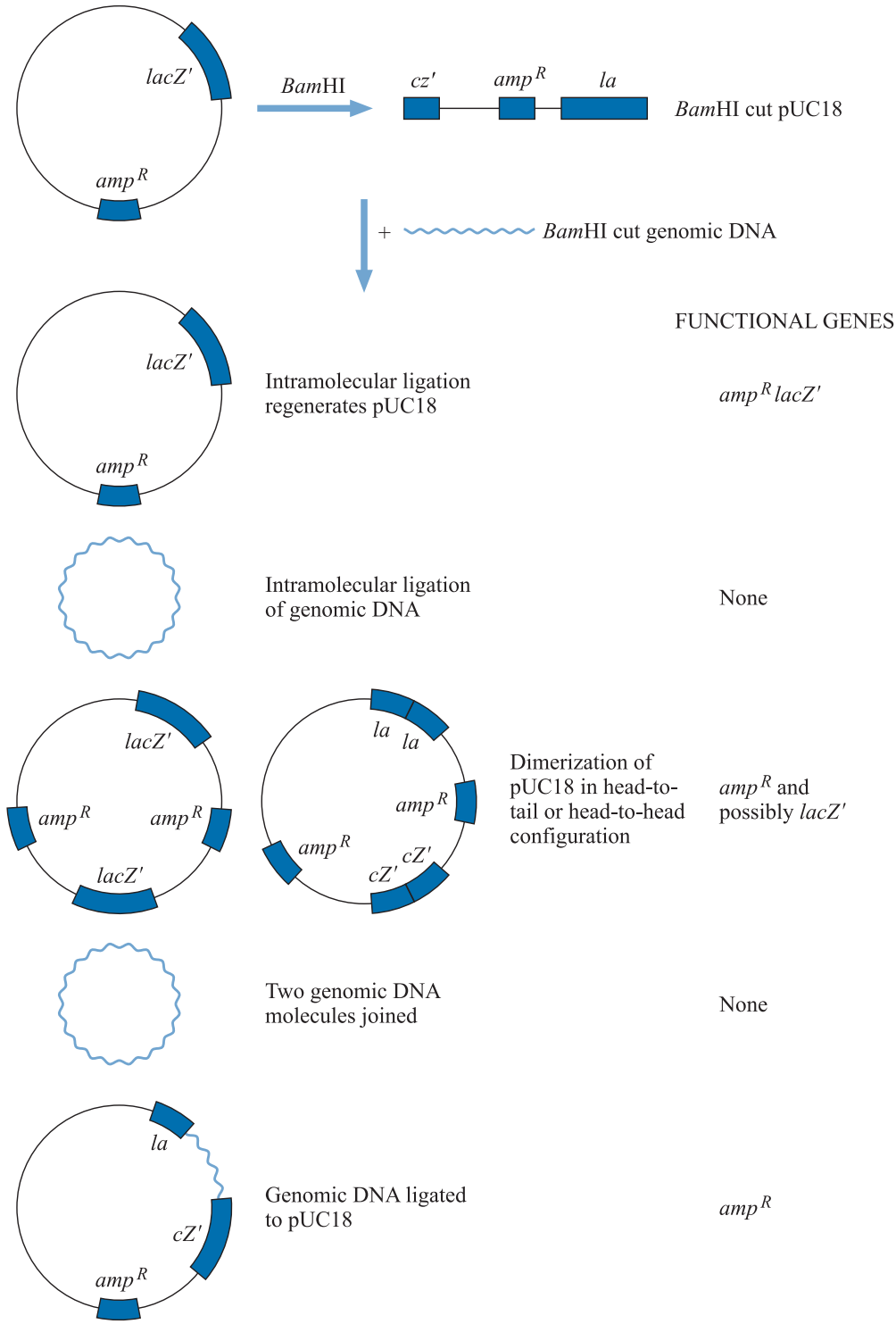


Fig 3.3 Ligation of *Bam*HI-cut genomic DNA into a *Bam*HI-cut vector. The figure shows the regeneration of vector, the circularization of genomic DNA and the ways in which two separate molecules (vector or genomic DNA) can be joined. The functional genes present on each molecule after ligation are indicated.

those cells that do are quite efficient at DNA uptake and may take up more than one piece. So a low, but significant fraction of colonies may contain two different plasmids. Over time, these may segregate, with some cells containing one plasmid and the rest containing the other.)

The problem now is to distinguish the colonies that have acquired genomic DNA from all the others. That can be done in this example by using the *lacZ′* gene. Colonies that have acquired pUC18 molecules without an insert of genomic DNA will have an intact *lacZ′* gene (which will be induced by the IPTG). They will, therefore, be able to break down the X-gal and be coloured blue. The colonies that have acquired molecules with a genomic DNA insert will have a *lacZ′* gene that has been inactivated by the insertion. Thus, these colonies will be unable to metabolize the X-gal and will be the usual white (actually off-white) colour. The head-to-tail dimeric pUC18 molecules will also give blue colonies, and the head-to-head dimers will not replicate stably in the host cell. The relative abundances of blue and white colonies will depend on factors such as the relative proportions of genomic and plasmid DNA in the initial ligation. We are interested in the white colonies. We can confirm that plasmids in the white colonies contain an insert by growing up the bacteria, purifying plasmid DNA and cutting it with a restriction enzyme. *Bam*HI would be the easiest in this case, and should generate two sizes of molecule: plasmid and inserted genomic DNA. Plasmids from blue colonies, where there is no insert, will give a single size of restriction fragment, corresponding to the plasmid DNA. The same would be seen for the head-to-tail dimeric plasmids. (An alternative way to differentiate whether plasmids have an insert, and to measure the size of the insert, is to do PCR on DNA from the colonies using primers that anneal on each side of the cloning site. This can be done directly with cells, without needing to purify the DNA, and is called colony PCR.)

The outcome of this experiment is a large number of *E. coli* colonies, each carrying a plasmid containing a *Bam*HI fragment of genomic DNA. This is useful in a number of ways. For example, to get a lot more of each piece of genomic DNA, it is just a matter of growing up a larger quantity of *E. coli*, extracting the plasmid and, if necessary, separating the insert from the plasmid. That is much easier than extracting the DNA from more of the organism(s) we used in the first place. Also, each plasmid insert is a single, defined piece of DNA. To separate a single, defined piece out of the total genomic DNA each time one wanted it would be too laborious. This propagation of individual, defined pieces of DNA in a suitable organism is cloning. The generation of recombinants in this way by randomly cloning fragments of the total DNA from an organism is called shotgun cloning, and the random collection of *E. coli* colonies containing the recombinant DNA molecules is called a library. Because total genomic DNA was used, it is called a genomic library. An ideal genomic library would contain all the sequences present

in the original organism's genome, and would then be described as representative. A library constructed in the way we have just described would probably not be very representative. For example, small DNA fragments tend to be cloned more efficiently than large ones, so our library would be underrepresentative of the large fragments. Another problem with the library in our example is that any gene that contained a *Bam*HI site could never be rescued intact, because *Bam*HI digestion was used to generate the fragments to be cloned. It would be better to use some other way of generating the fragments (see Section 5.2).

3.1.2 Use of the *lacZ′* indicator system (alpha-complementation)

In this example, we said that the pUC plasmid contained only a part of the *lacZ* beta-galactosidase gene. Beta-galactosidase is rather a large protein (116 kDa), and it is more convenient to have just a portion of the gene on the plasmid to keep the plasmid as small as possible. This portion encodes the first 146 amino acids and is sometimes called the *lacZ′* minigene. In addition to the *lacZ′* gene and the ampicillin resistance gene, the plasmid also contains a *lacI* gene for the Lac repressor protein (Figure 3.1). This keeps the minigene repressed, except in the presence of an inducer such as IPTG. This may be particularly important if the sequences inserted into the minigene are toxic if expressed at high levels. The rest of the *lacZ* gene is contained within the host, often on an F′ plasmid. In fact it is a version of the gene called the M15 deletion, which lacks codons 11–41. The polypeptide produced from this fails to tetramerize, and this tetramerization is needed for enzymatic activity. However, in the presence of the minigene product (amino acids 1–146), assembly can take place to produce a molecule with low but detectable beta-galactosidase activity. This intragenic complementation is called alpha-complementation. The principle of alpha-complementation is summarized in Table 3.1. In essence, it is just a way of reducing the size of the vector needed. The pUC vectors were developed at the University of California (hence their name). We will now go back to the basic procedure we have described and look in a little more detail at the vectors and hosts.

Table 3.1 Alpha-complementation

	LacZ residues 1–146 present?	M15-deleted LacZ present?	Result (+IPTG+X-gal+Amp)?
Host	No	Yes	Amp^S, no LacZ; no colonies
Host + pUC plasmid	Yes	Yes	Amp^R, LacZ; blue colonies
Host + pUC plasmid (with insert)	No	Yes	Amp^R, no LacZ; white colonies

3.2 Vectors, transformation and hosts

3.2.1 Vectors

As outlined above, vector is the name given to DNA molecules into which foreign DNA is inserted for subsequent propagation in a host cell. A vector should have several features:

1. **An origin of replication**. Without an origin of replication, the vector could not replicate, and when the cells divide after taking up the vector molecule, only one of the daughter cells would retain it. Therefore, we would never get a colony of transformed cells. The pUC18 origin of replication came ultimately from a plasmid in a clinical bacterial isolate denoted as pMB1.
2. **A selectable marker**. This is needed to distinguish cells that have taken up the vector from those that have not. In the example we have just looked at, the selectable marker was ampicillin resistance. The ampicillin resistance gene derives from a transposon (a piece of DNA able to move, or transpose, to different places in the genome) from another plasmid, pRSF2124.
3. **Suitable single restriction sites.** In the example we considered, pUC18 had just one recognition site for the enzyme *Bam*HI, which was used in the cloning. Had there been more than one *Bam*HI site in pUC18, the vector would not have been suitable, because cutting with *Bam*HI would have cut it into more than one piece. Reassembly of the vector together with an insert during the ligation would have required a trimolecular reaction at least, which would have been very unlikely. Note, though, that the same enzyme does not necessarily have to be used for cutting the vector and insert (see Section 1.2.1). If the insert molecules had been generated by *Sau*3A digestion, we could still have used *Bam*HI to cut the vector, as *Bam*HI ends are compatible with *Sau*3A ends. Cutting the vector with *Sau3*A would not have been feasible, as there are too many sites for *Sau3*A in the vector. Note also that restriction sites located in indispensable genes in the vector are unsuitable for cloning, because insertion of DNA there would be likely to destroy gene function and, therefore, vector viability. We will see that this is particularly important when we consider vectors based on bacterial viruses in Chapter 4.
4. **Suitable size.** To some extent, having a suitable size is a corollary of having suitable single restriction sites. A restriction enzyme cleavage site that comprises six nucleotides will occur on average approximately once every 4^6 bp (i.e. every 4 kbp or so). So a vector that was much larger than 4 kbp might be expected to have several sites for a given enzyme, and cutting the vector would reduce it into several pieces that would be unlikely to be correctly assembled in a ligation reaction. It is possible to remove restriction sites, and methods for doing this are described in Chapter 7. Simply removing excess restriction sites may not eliminate

the problem, though. Large DNA molecules are very susceptible to physical shearing, even in the simple act of pipetting, so they are always difficult to handle. Note that the use of the alpha-complementation system allowed the size of the pUC plasmids to be reduced from what would otherwise be necessary to encode beta-galactosidase.

5. **Markers for DNA insertion.** In the example we studied, the insertion of DNA into the vector can be detected by inactivation of the *lacZ′* gene, which in turn could easily be assayed by plating cells onto medium containing an inducer and a chromogenic substrate as well as ampicillin. In some older vectors, an additional round of plating was needed to distinguish recombinants from non-recombinants. An example of this is the vector pBR322, which contains an ampicillin resistance gene and a tetracycline resistance gene (Figure 3.4). There are cloning sites in both genes. Suppose we had used the *Bam*HI site in the tetracycline resistance gene for insertion of DNA. After transformation, we would plate the *E. coli* cells on medium containing ampicillin to select for the acquisition of a plasmid. We would then plate small samples from ampicillin-resistant colonies onto medium containing tetracycline. If DNA had been inserted into the *Bam*HI site, then the tetracycline resistance gene would be inactivated, and cells containing recombinant plasmids would be sensitive to tetracycline. Cells that had acquired a plasmid without inserted DNA would be resistant to tetracycline as well as ampicillin. This need for two rounds of plating was one of the reasons for the pBR322 plasmid being superseded by the pUC and other related plasmids as a routine cloning vector. In some of the first plasmids used for cloning, the cloning site was not in a functional sequence at all. The only way to detect insertion, therefore, was to isolate plasmid DNAs, digest with a suitable enzyme and examine the

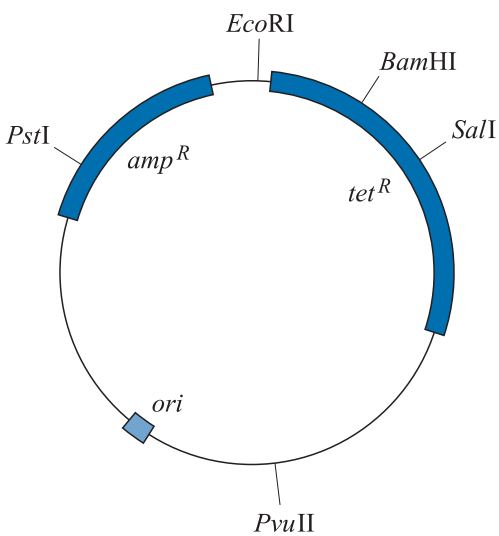


Fig 3.4 The plasmid pBR322 (4.3 kb). The figure shows genes for ampicillin resistance (*amp*^{R}), tetracycline resistance, (*tet*^{R}), the origin of replication (*ori*), and a selection of unique restriction sites.

fragments produced electrophoretically. That was potentially very tedious, especially if intramolecular ligation had been favoured over the intermolecular reactions (the tendency for which depends on the relative concentrations of plasmid and genomic DNA), in which case most of the plasmids in transformed cells would not be recombinant.

6. **High copy number.** Having a high copy number is desirable but not essential, like having markers for DNA insertion. To maximize the yield of plasmid from transformed cells, the copy number in each cell should be as high as possible. Different plasmids have different copy numbers. For some, such as the F factor, the copy number is low, perhaps one or two per cell. The replication of the F factor is quite closely tied to replication of the chromosome, and such control of replication is said to be stringent. The plasmid pSC101 (one of the first plasmids used for cloning) is also under stringent control and is usually present at no more than five copies per cell. Other plasmids with different origins of replication have less tightly controlled replication, and they are said to be relaxed. Here, the copy number can range widely. The copy number of pBR322 is typically 15–20 per cell, whereas vectors such as the pUC plasmids may be present in 500–700 copies per cell. The origin of replication for both pBR322 and the pUC plasmids came ultimately from the plasmid pMB1. The reason for the difference in copy number between these plasmids with, in theory, the same replication origin may lie in a mutation in the pUC plasmids in the region encoding the RNA molecules that regulate replication by interacting with the origin. This mutation makes the repression system that controls plasmid replication less effective. The copy number of many low copy-number plasmids containing the pMB1 origin can be increased by chloramphenicol amplification. The host cell culture is treated with chloramphenicol, which inhibits bacterial protein synthesis. The inhibition of protein synthesis blocks chromosomal DNA replication, because particular proteins, such as DnaA, need to be synthesized each time chromosomal replication is initiated. The inhibition also blocks cell division, which is closely tied to chromosomal DNA replication. Plasmid replication, however, requires only proteins that are more long-lived. It can, therefore, continue when chromosomal DNA replication and cell division have stopped. Eventually, plasmid DNA replication will stop too, as the supply of general replication proteins, such as DNA polymerase, runs out; but the average copy number will have increased greatly.
7. **Disablement.** Ever since the earliest experiments in genetic manipulation, there has been concern over the possibility of recombinant DNA molecules 'escaping' into the environment and spreading. The likelihood of this can be reduced if the plasmid is in some way disabled so that it cannot spread to other bacteria by processes such as conjugation. Many plasmids, such as pBR322,

have been disabled by removal of the *mob* gene, which is required for them to mobilize themselves by conjugation. However, such plasmids can still be transmitted from cells containing other plasmids that can provide the necessary functions for mobilization. This transmission can be blocked by removing a region containing sites called nic and *bom* from the plasmid that might be mobilized. This region is where the proteins provided by the other plasmids act. The pUC vectors are among those that have had this region removed.

3.2.2 Transformation

The transformation method chosen will depend on the efficiency required. The more complex chemical methods for inducing competence described in Section 1.5.2 can generate up to 10^9 transformants per microgram of plasmid DNA, with up to 5% of viable cells competent for transformation. If the recombinant molecules are abundant or of just one (or a few) types, as when reintroducing a plasmid DNA stock into cells to prepare more of the same plasmid, then simple treatment with calcium chloride will be sufficient. If the number of recombinant molecules available is very small and as many different members of the collection as possible need to be recovered (e.g. in cDNA cloning, see Section 5.3), then a more efficient system such as electroporation may be required. When optimized, this approach can yield a higher number of transformants per microgram of DNA than any of the chemical treatments (up to 5×10^{10} per microgram have been claimed). Typically, a field strength of the order of 15 kV cm^{-1} is used, with a decay time constant of about 5 ms. These conditions will usually result in most of the cells remaining viable, and a reasonable balance between cell killing and the induction of competence.

3.2.3 Hosts

The host is the cell in which the recombinant molecules are to be propagated. Choosing the right host is as important as choosing the right vector. Essential or desirable characters include the following:

1. **Efficient transformation.** This depends on two main features. One is the ability actually to take DNA into the cell, and this process is poorly understood. Different genotypes respond differently to different transformation systems. Some mutations appear to enhance the efficiency of transformation itself. These include the *deoR* mutation, which seems to assist particularly in the uptake of larger DNA molecules. DeoR is a transcriptional regulator with DNA-binding activity that controls expression of a set of genes involved in deoxyribonucleoside catabolism.

   The main feature determining transformation efficiency is the presence or absence of endogenous DNA-degrading systems.

Many hosts used for cloning are derived from *E. coli* strain K, which contains the K restriction–modification system encoded by the *hsdRMS* locus. The *hsdR* gene encodes an endonuclease that cleaves DNA containing the sequence -AACNNNNNNGTGC-, unless the second of the two adenine residues and the adenine residue on the other strand opposite the thymine are methylated. Many hosts, therefore, have an *hsdR* mutation (or a larger deletion) to avoid cleavage of incoming unprotected DNA. The *hsdM* gene encodes the methylase that protects against degradation so passaging of DNA through an $hsdR^{-}M^{+}$ strain can be used to allow methylation if it is subsequently necessary to propagate in an $hsdR^{+}$ strain.

There are other proteins in *E. coli* strain K that will degrade incoming DNA if it *is* methylated, belonging to the methylation-dependent restriction systems (MDRS). These are not as well understood as the Classes I–III restriction–modification systems. They include the endonucleases that are the products of the *mcrA*, *mcrB* and *mrr* loci, which will degrade DNA containing methylcytosine (*mcrA*, *mcrB*) or methyladenine (*mrr*). Using strains mutant in these loci, therefore, is desirable, particularly when cloning highly methylated DNA.

The extent of methylation of DNA in the host may also affect the efficiency with which other restriction enzymes will subsequently cleave the DNA in vitro. Methylation can be carried out by the enzymes mentioned already, and also by the products of the *dam* and *dcm* genes. The Dam protein methylates adenines at the sequence -GATC-, and Dcm methylates cytosines at the sequences -CCAGG- and -CCTGG-. Some frequently used restriction enzymes have recognition sites that overlap with these and are inhibited by methylation, so DNA prepared from strains that are wild type for these loci will not be efficiently restricted. Use of dam^{-} or dcm^{-} strains has the disadvantage, however, that they have high mutation rates. This is because newly replicated dsDNA is hemi-methylated (as the newly synthesized material has not yet been methylated). If the polymerase has introduced any errors during synthesis, they will result in a mismatch; these mismatches are normally resolved in the cell by correction to the *methylated* strand. In a dam^{-} or similar strain, neither strand is methylated, and the mismatch correction is as likely to be to the incorrect (newly synthesized) strand as to the correct one.

2. **Stable maintenance of plasmid.** Once a recombinant plasmid has entered the cell (assuming it has escaped any endogenous restriction enzyme activity), it is still not guaranteed to replicate indefinitely and stably even if it has a suitable origin of replication. Rearrangement of the recombinant may occur, and the most frequent manifestation of this is partial deletion (i.e. loss of part of the molecule). This usually occurs by recombination across directly repeated sequences, as shown in Figure 3.5. Not all the plasmid molecules in one cell need undergo the same

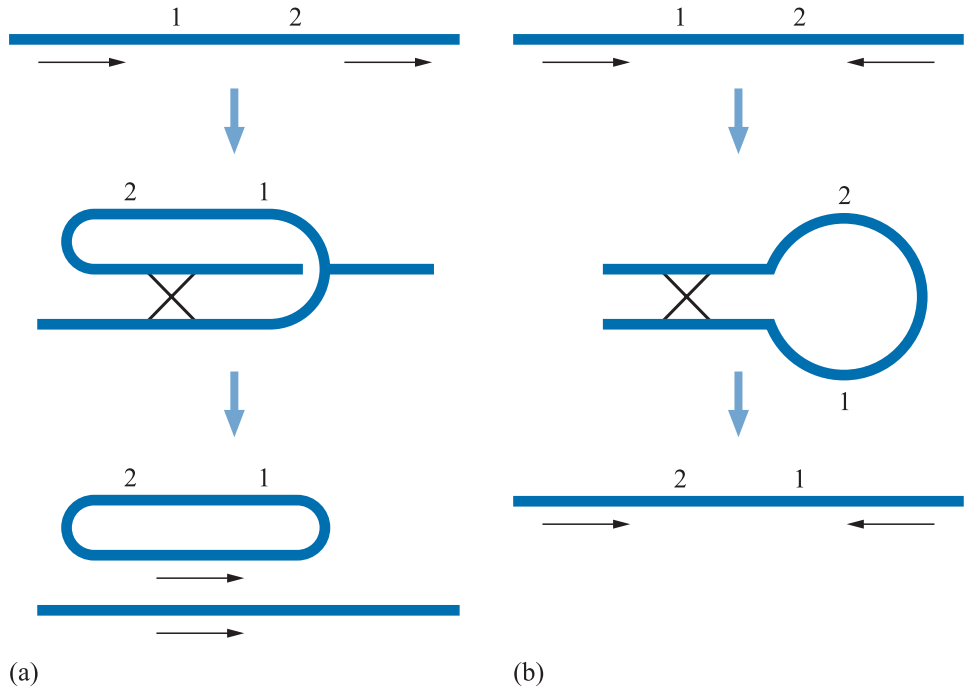


Fig 3.5 Recombination across repeated sequences. If the repeats are in a direct configuration, then the material between them is lost (a). If they are inverted, then the material between them is inverted on recombination (b). 1 and 2 represent arbitrary points on the molecule.

recombination at the same time, so recombination across a pair of directly repeated sequences in a plasmid will leave three types of molecule in the cell: the original, unrecombined one; and the two smaller recombination products. One of the products will lack an origin of replication and be lost from the population. The other will still be able to replicate, and will do so faster than the original molecule because it is smaller. Over a number of rounds of replication, even a small difference in the time required for replication can result in a large difference in the number of molecules present. Generating 1 μg of DNA from a single copy of a 4 kbp plasmid takes about 40 cycles of replication. A plasmid that could replicate 10% faster would, over this period, generate approximately 16 times as much DNA and, therefore, be by far the most abundant molecule in the total population. The problem will be exacerbated if the original plasmid is deleterious to the host, perhaps by expression of a protein that is toxic to the cell, and the partial deletion caused by recombination abolishes production of that protein. The possibility of preferential propagation of altered molecules during cloning is, therefore, one that should always be borne in mind. Recombination need not always result in deletion. It can also cause inversion, if it takes place across inverted repeats (Figure 3.5).

Since deletions and inversions usually depend upon recombination, mutations in the host that suppress recombination will help to ensure the stability of transforming molecules. There are three main recombination systems in *E. coli*, using the products of the *recBCD*, *recE* and *recF* genes. However, these systems depend largely on the product of the *recA* gene, so strains mutant in this will have a greatly reduced recombination frequency. Some sequences, particularly those containing inverted repeats, may be subject to rearrangement in a *recA*-independent way. This may be due to the *recF* pathway functioning in the absence of (or low levels of) *recA*, so an *recF* mutation may be desirable in addition to *recA*. It is also reported that deficiency of the DNA gyrase encoded by the *gyrA* and *gyrB* genes can lead to enhanced stability of molecules containing repeated sequences.

3. **Disablement.** As with vectors, it is necessary to take precautions to ensure that strains carrying recombinant plasmids are unlikely to escape and propagate outside the laboratory (i.e. to enhance containment). For this reason, the preferred strains usually carry mutations that reduce their viability in the wild. These are often mutations conferring auxotrophy (i.e. the requirement for a particular metabolite to be supplied in the medium, resulting from an inability to synthesize the metabolite).
4. **Features allowing use of the lacZ′ minigene**. We saw earlier that the use of the *lacZ′* minigene requires the use of a host producing a suitable LacZ fragment, and that this is sometimes encoded on an F′ plasmid. In these cases, therefore, it is necessary to be able to ensure retention of the F′ plasmid. For this reason, the F′ plasmid also contains genes *proA* and *proB* for two of the enzymes of proline biosynthesis (encoding 5-glutamyl phosphate reductase and glutamate-5-kinase respectively), and these are deleted from the chromosome of the host. Propagation of the host on medium lacking proline ensures that only cells carrying the resident F′ plasmid will grow.
5. **Other markers.** In any work with bacteria, it is very useful to have strains that are genetically marked, so the correct strains can be recognized as such. The recombination and nutritional deficiency markers mentioned above may be useful in this respect, although it may also be convenient to have other markers that can be selected more easily. Antibiotic resistance can be useful, although of course it should be different from any of the antibiotic resistances conferred by the vectors to be used. Another mutation that is often incorporated into hosts is *endA*, inactivating the gene for a DNA-specific endonuclease. This mutation enhances both the yield and quality of plasmid DNA preparations. The *relA* mutation overrides controls on RNA synthesis, improving the rate of synthesis in the absence of protein synthesis. The *tonA* mutation causes loss of an outer membrane protein, to which phages T1 and T5 adsorb. Infection by these phages can be a problem in the laboratory, so the *tonA* mutation is useful in conferring resistance

to phage infection. Host strains used for particular purposes, such as the expression of cloned genes, have additional features, as discussed in Chapter 8.

It may be helpful to look at the genotype of a common host strain, and see how it fulfils the considerations above. The strain DH5α is a common host for the pUC plasmids, and is a derivative of *E. coli* strain K. Its genotype is *endA1 hsdR17 supE44 thi-1 recA1 gyrA (nal^R) relA1 Δ(lacZYA-argF)U169 Φ80lacZΔM15*

endA1 is discussed above.

hsdR17 inactivates the host restriction system, allowing DNA to be protected by methylation, but not digested, which is sometimes abbreviated to $r_K^- m_K^+$.

supE44 is an 'amber' chain termination suppressor mutation that allows readthrough of UAG codons in translation. This is used because the same host (or derivatives of it) can also be used for propagation of phages, a number of which have amber chain termination mutations as a biological containment measure. Growth of such phages is possible only in a suppressor host.

thi-1 is a nutritional requirement (for thiamine) that gives some containment.

recA1 gyrA (nal^R) relA1 are discussed above. (*nal^R*) is resistance to the DNA gyrase inhibitor nalidixic acid, conferred by the *gyrA* mutation.

Δ(lacZYA-argF)U169 is a deletion of the *lac* operon (extending into an arginine biosynthesis gene) that is required for the blue-white selection.

Φ80lacZΔM15 describes a ***Φ80lacZ*** prophage that directs synthesis of the partially deleted LacZ, required for blue–white selection.

Another example of a widely used host is JM109, also a derivative of *E. coli* strain K. It is *endA1 hsdR17 supE44 thi-1 recA1 gyrA (nal^R) relA1 mcrA Δ(lacZYA-proAB) F′ traD36 proA$^+$ proB$^+$ lacI^q lacZΔM15*.

Many of the mutations are similar to those in DH5α. JM109 also carries the ***mcrA*** mutation, described earlier, and a different *lac* deletion, extending into the proline biosynthesis operon. The symbol *F′* indicates that the markers following it are on an F′ plasmid.

traD36 reduces the efficiency of conjugation by a factor of 10^5, contributing to biological containment.

proA$^+$proB$^+$ restores the ability to synthesize proline that was lost by the deletion into *proAB*. This means that growth of the strain on a medium that lacks added proline selects for the presence of F′, as it requires the cells to synthesize their own proline.

lacI^q is a mutant form of the Lac repressor gene, causing increased levels of repressor in the cell, and allowing tighter control of the *lacZ* gene. This is important for the use of this host for vectors that replicate to a high copy number. Such vectors might titrate out all the repressor produced by a wild-type *lacI* gene, especially if the gene was present only on the chromosome or on a low-copy-number plasmid

such as F′. Titration of the repressor might lead to uncontrolled transcription of sequences inserted within the *lacZ* region of the vector.

lacZΔM15 is the partially deleted *lacZ* needed for alpha-complementation.

Notice that JM109 allows selection for retention of the F′ plasmid, as described. This is important for two reasons. The partially deleted *lacZ* needed for alpha-complementation is encoded on the F′. Also, the host is sometimes used for growing phages (Section 4.3) that propagate only in a host carrying the F (or F′) plasmid.

3.3 Modifications

3.3.1 Avoiding self-ligation

Although the blue–white screening that is available with the pUC vectors allows us to see which colonies on a plate contain recombinant plasmids, these are sometimes very few in number compared with the colonies containing vector that religated without any insert. If we need to recover a large number of different recombinant molecules, then the large number of non-recombinant colonies may be a problem. There are various approaches to minimizing this. One approach is to increase the ratio of insert DNA to vector DNA. A more reliable solution is to use alkaline phosphatase (Section 1.3.1). This enzyme is capable of removing 5′-terminal phosphate groups from nucleic acid molecules (and nucleotides). However, these phosphate groups are required for ligation to take place. So if the vector is treated with alkaline phosphatase to remove its terminal phosphate groups, then self-ligation is no longer possible. Ligation of vector to insert, however, is still possible because each insert molecule carries two phosphate groups, one at each end. Note though, that sealing each vector–insert boundary requires two phosphate groups, one for each strand. Phosphatase treatment of the vector, therefore, means that each boundary can be covalently sealed by ligase on one strand only. The other strand will remain nicked, as shown in Figure 3.6. However, these nicks are not a serious problem. Transformation is still possible, and the normal cellular replication and repair processes will produce molecules that are no longer nicked.

In practical terms, the alkaline phosphatase treatment is carried out after the vector has been cut, and before vector and insert DNA are mixed. All traces of phosphatase activity must be removed before the insert DNA is added, otherwise the insert would be dephosphorylated, too, and intermolecular ligation will also be blocked. Commonly used alkaline phosphatases are inactivated by heating for a few minutes. This treatment can be followed by removal of protein with phenol and chloroform to minimize any carry over of active enzyme into the ligation reactions. If the enzyme used is sufficiently heat labile, purification of the DNA prior to ligation can be omitted.

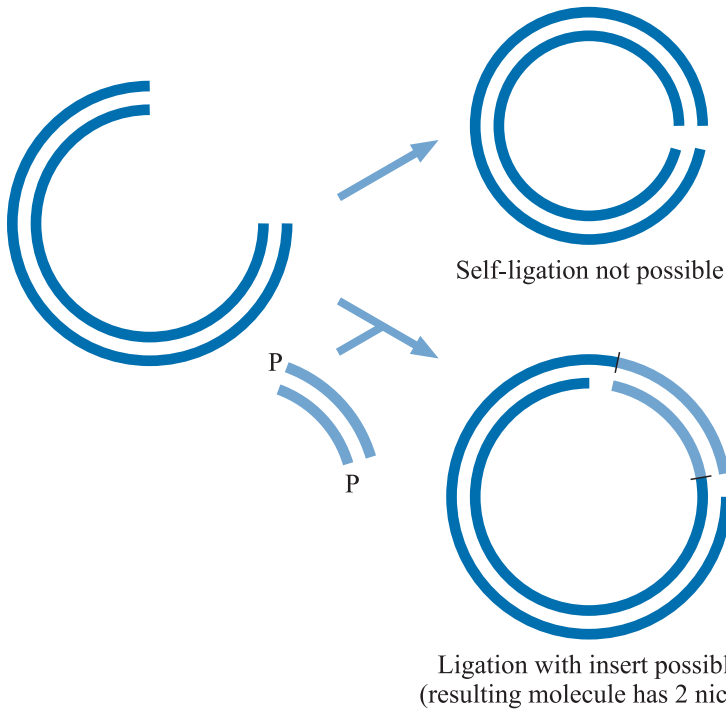


Fig 3.6 Phosphatase treatment of vector. Removal of the phosphate groups (P) from the vector means that self-ligation is not possible. Ligation to a molecule with 5′-phosphates is still possible, although one strand will be nicked at each junction.

A third approach to increasing the fraction of recombinants recovered is to use more complex vectors that, if self-ligated, direct the synthesis of a molecule that kills the host cell. One example uses the *ccdB* (control of cell death) gene. The coding region for this is arranged in the vector so that self-ligation results in the expression of the CcdB protein and consequent death of cells containing the molecule. This is the basis of 'Zero Background™' vectors.

3.3.2 Forced cloning

In the example we considered, the insert DNA could have been ligated into the vector DNA in either of two orientations. In some circumstances it is desirable to control the orientation in which a particular restriction fragment is inserted. This is sometimes called forced cloning. This is possible with the pUC vectors, as there are many different restriction sites clustered in the multiple cloning site. The region is also sometimes called the polylinker. The consequence of having this collection of sites is that, if the insert DNA has different restriction sites at each end, we can readily use a vector cut with the two corresponding enzymes. There will then be only a single orientation in which the insert can be ligated to the vector.

There is a series of pUC vectors with different combinations and arrangements of restriction sites in the multiple cloning site. They generally come in pairs whose multiple cloning sites are mirror images of one another, as shown in Figure 3.7. They were generated

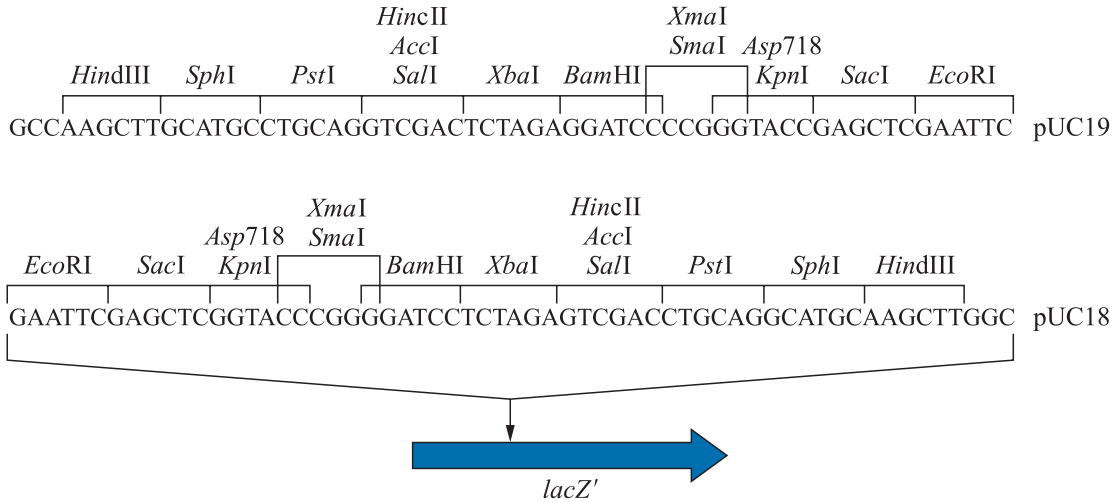


Fig 3.7 Multiple cloning sites of pUCI9 and pUCI8. Note that they contain the same restriction sites but in the opposite orientation. So an *Eco*RI—*Hin*dIII fragment, say, could be cloned into either vector, but would have opposite orientations.

by the ligation of different synthetic DNA molecules containing the appropriate restriction sites into an *Eco*RI site close to the beginning of the *lacZ'* minigene.

This wide combination of restriction sites, combined with the blue—white screening, makes the pUC vectors very versatile. There are many other series of vectors based on the same principle, but with additional features (e.g. to drive transcription of the insert DNA). We shall deal with some examples of these later.

3.3.3 Cloning PCR products

The basic cloning method can be applied to the cloning of PCR products. In the simplest case, we can treat the PCR products as blunt-ended DNA molecules and ligate them into a vector that has been cut to give blunt ends. Very often, this blunt-end ligation is not very efficient. A better approach is to incorporate restriction sites into the primers used for PCR. The PCR products will, therefore, have these restriction sites located at each end and they can be cut with the appropriate enzyme(s) prior to a sticky-ended ligation into an appropriately cut vector. It can be even more efficient to exploit the fact that many polymerase preparations used for PCR incorporate an additional A residue that is not template encoded onto the end of the molecules they synthesize. The products can, therefore, be ligated directly into vector molecules that have an overhanging T residue. A suitable vector DNA preparation can be made by cutting vector with an enzyme that generates blunt ends and then enzymatically adding a 3′ T residue to the ends. This is the basis of the pGEM-T vector series, which are available commercially as linearized plasmids with the T residue added.

3.3.4 Methods that do not use DNA ligase

In the basic experiment, we used restriction enzymes to cut DNA molecules and ligase to rejoin them. There are other systems available that have different ways of cutting and ligating molecules. They are often faster than approaches based on DNA ligase, but they are newer and less widely used.

1. **Topoisomerase.** This approach uses the *Vaccinia* virus topoisomerase I. In vivo, it cleaves one strand of a DNA molecule at the recognition sequence -C/TCCTT- and becomes covalently attached to one end of the cleaved molecule. The supercoiling of the DNA template can then be altered by swivelling around the uncleaved strand, and the enzyme then reseals the original strand and dissociates from the DNA. Commercially available topoisomerase-based vectors ('TOPO® Cloning' systems) are supplied in linear form, with the topoisomerase enzyme attached. The topoisomerase will rapidly ligate the vector to suitable substrate molecules. For cloning PCR products with overhanging A residues, the vector has a single overhanging T residue with the topoisomerase attached. The PCR product can anneal to the vector and similarly be rapidly ligated.
2. **Recombinase.** Site-specific recombination reactions can be used to clone PCR products or transfer inserted pieces of DNA rapidly from one vector to another. This is the basis of the 'Gateway®' system, based on bacteriophage lambda (Section 4.4). This bacteriophage can integrate its DNA by site-specific recombination between a site on the phage (*attP*) and a site on the host genome (*attB*). Recombination across *att* sites is exploited in the Gateway system to transfer inserts directly between vectors carrying *att* sites. For cloning PCR products, for example, the primers contain the *attB* site. The recipient vector (called an entry vector) contains *attP* sequences flanking the insertion site. Addition of recombinase protein catalyses recombination between *attB* and *attP* sites, integrating the PCR product into the vector, as shown in Figure 3.8. The insert can then be transferred, if required, into different vectors that have suitable *att* sites by similar site-specific recombination. Alternatively, restriction fragments can be inserted into an entry vector by conventional ligation and then transferred into other vectors by recombination.

3.4 Linkers, adaptors and cassettes

3.4.1 Linkers

The importance of the pUC and similar vectors owes a great deal to the presence of the multiple cloning site, with its ability to accept a wide range of different fragments. A similar feature is offered by

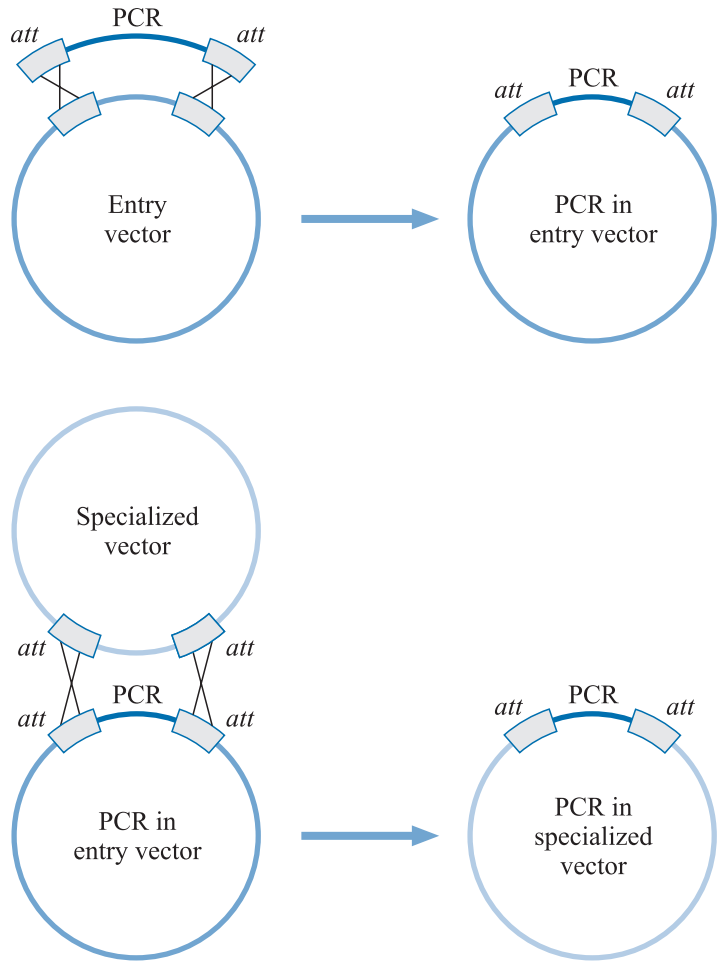


Fig 3.8 'GATEWAY®' cloning system. In the first stage a PCR product (PCR) is transferred into an entry vector by site-specific recombination across *att* sites. In the second stage, the product is transferred into a specialized vector by recombination.

molecules called linkers. These are short, chemically synthesized molecules that contain a particular restriction enzyme recognition site within their sequence. An example of these, an *Eco*RI linker, is shown in Figure 3.9. Using *Eco*RI linkers, it is possible to clone a blunt-ended insert molecule into a vector carrying an *Eco*RI cloning site. The linkers are themselves blunt-ended molecules, and can be joined onto the blunt-ended insert molecule using T4 DNA ligase. Although the reaction (being a blunt-ended ligation) is relatively inefficient, the use of an excess of linker helps to ensure that a large proportion of the insert molecules have linkers on the ends. Some molecules end up with more than one linker attached to each end, but that is not a problem. The next step is to treat the insert-linker molecules with the appropriate restriction enzyme, which cuts within the linkers to leave a single cut linker attached to each end of the insert. So the insert now has sticky ends, which can be used for insertion into a restriction site in the usual way. This is summarized in Figure 3.9.

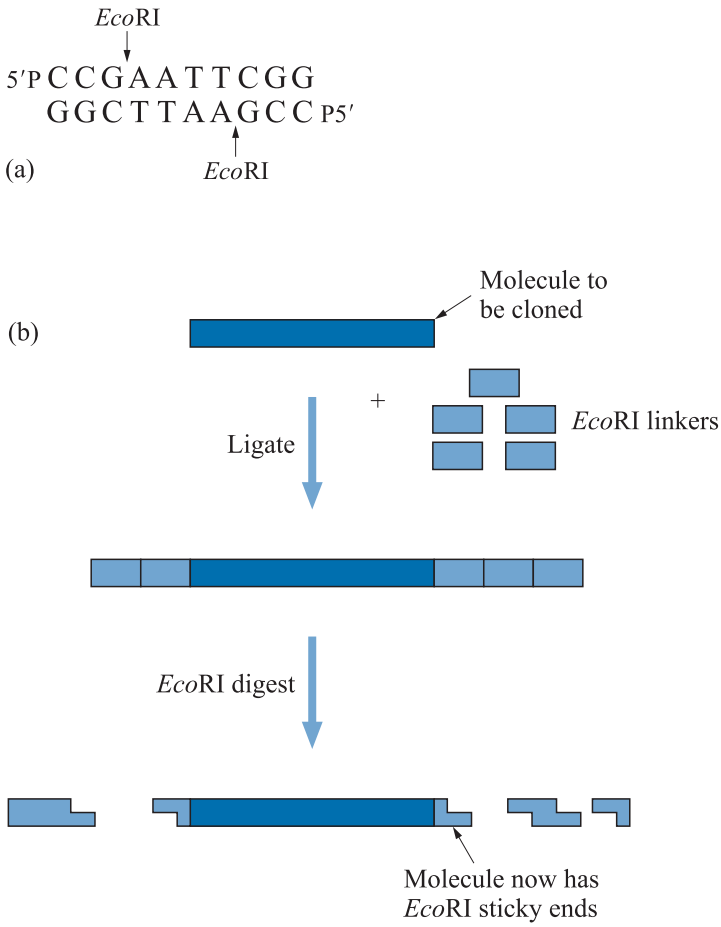


Fig 3.9 An *Eco*RI linker. An example of a linker is shown in (a). The use of linkers is shown in (b). The linkers are first ligated onto a target molecule in a blunt-ended reaction and are then cleaved with *Eco*RI.

There is a potential problem, though. If the insert itself contains a recognition site for the linker restriction enzyme, then the second step of the process in Figure 3.9 will cut the insert as well as the linkers. This is likely to be undesirable. It can be avoided by treatment of the insert, prior to the addition of linkers, with the appropriate methylase: *Eco*RI methylase if *Eco*RI linkers are being used. This renders the molecule insensitive to the restriction enzyme, which will still be able to cut the linker.

Of course, an alternative way of cloning blunt-ended DNA into a vector with *Eco*RI ends is to avoid using linkers and instead polish up the ends of the cut vector to make them blunt. That could be followed by a blunt-ended ligation of insert DNA into vector. What are the advantages of using linkers? There are three main advantages. One is that it makes better use of the insert and vector DNA, which may be in short supply. Ligation of blunt-ended insert into blunt-ended vector is an inefficient reaction, with much of the insert being wasted. Ligation of blunt-ended linker onto blunt-ended insert can also be inefficient. However, we use a large excess of linkers

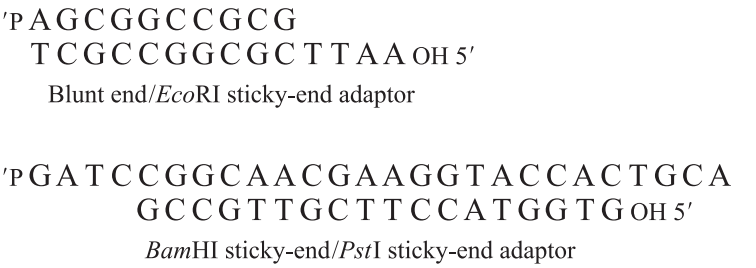


Fig 3.10 Adaptors. The upper molecule has a blunt end (left-hand side) and an *Eco*RI sticky end. It can be used to convert a blunt-ended molecule into an *Eco*RI-ended molecule. The lower molecule has a *Bam*HI end (left-hand side) and a *Pst*I end (right-hand side). Therefore, it can be ligated onto a molecule with *Bam*HI ends to produce one that has *Pst*I ends (or vice versa, depending on the availability of phosphates on the target molecule).

(which are in abundant supply, being synthesized on a chemical scale), so that a large proportion of insert acquires linkers and can be efficiently ligated into the vector. The second advantage, which also comes with using an excess of linkers, is that the likelihood of two insert DNA molecules being ligated to each other (concatenated) is reduced. The third advantage comes when it is necessary to cut the insert out of the vector at a later stage. It is unlikely that ligation of a blunt-ended insert into vector that has been cut with *Eco*RI and then polished will regenerate any restriction sites. It is difficult, therefore, to excise the insert precisely from the vector. However, insert cloned into *Eco*RI-cut vector using *Eco*RI linkers can readily be excised simply by redigestion with *Eco*RI.

There is no reason why a linker should have only one restriction site within it. Linkers that contain a number of sites are available. These are often called polylinkers, and resemble the multiple cloning site of the pUC vectors.

3.4.2 Adaptors

One or both ends of a linker may be single stranded. These are sometimes called adaptors. Examples are given in Figure 3.10. The first shown is an adaptor that is blunt at one end (like a conventional linker) and sticky at the other. Therefore, it can be ligated, without further digestion, to a blunt-ended molecule to leave a sticky end. Note that in this example the 5′ overhanging end is not phosphorylated; this lack of a phosphate group prevents the concatenation of adaptors in the ligation reaction (which would obscure the sticky ends to be used for cloning). Some adaptors (such as the second example in Figure 3.10) are sticky at both ends and can be attached to a molecule that is already sticky-ended. Adaptors may also have extra restriction sites within their sequence.

3.4.3 Cassettes (cloning cartridges)

Cassettes (cloning cartridges) are a combination of linkers with other features, such as selectable markers. In their simplest form they consist of an antibiotic resistance gene flanked by DNA that contains multiple cloning sites (Figure 3.11). Consequently, they can be used as an easy way of incorporating selectable markers or other features into DNA molecules. This can be particularly useful as a way of inactivating a cloned gene in the technique of gene disruption

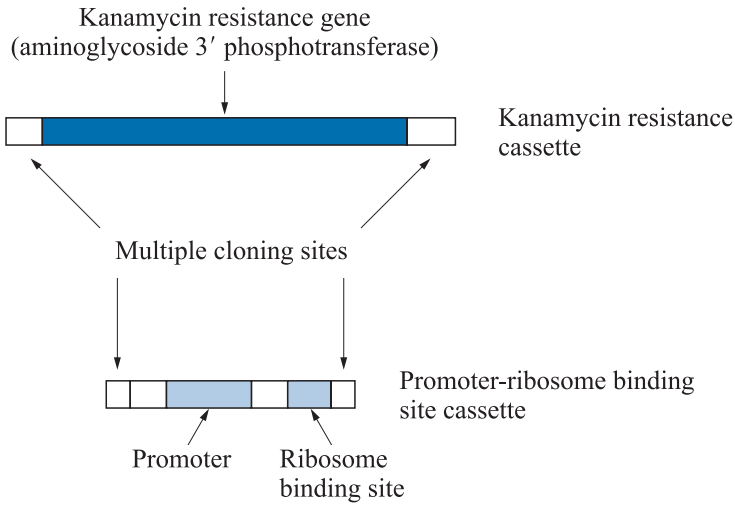


Fig 3.11 Cassettes. The diagram shows examples of an antibiotic resistance cassette (upper) and an expression cassette (lower).

(see Section 7.6). Cassettes may contain gene expression signals (such as promoters, terminators, etc.) rather than selectable markers. These cassettes are called promoter cassettes, terminator cassettes, and so on.

Chapter 8

Use of cloned DNA

8.1 Use as DNA

Cloned DNA can be used directly in too many different ways to describe here. They include sequencing, blotting (Southern, northern, southwestern, etc.), transcript mapping, footprinting and bandshift (gel retardation) assaying. Details of these can be found in general molecular and cell biology textbooks. Cloned DNA can also be used to build microarrays, as described in Chapter 1. We will concentrate first in this chapter on the use of cloned DNA sequences for expression, i.e. for directing the synthesis of RNA or protein. As we shall see, many cloning vectors have been developed specifically for this purpose. They are called expression vectors. We will then look at the use of cloned genes as tools for studying the function of other sequences.

8.2 Synthesis of RNA

8.2.1 Why synthesize RNA?

It may be necessary to produce RNA for a number of reasons. We might be studying RNA processing events, such as splicing or cleavage, in vitro and, therefore, need to produce RNA of a single type in the presence of as few contaminating proteins as possible. Or the aim might be to make a protein in a radiolabelled form by translation in vitro of an appropriate RNA in the presence of radioactive amino acids.

8.2.2 Vector systems

One approach to making RNA might be to use a vector with a powerful promoter to direct transcription in a bacterial cell in vivo and then isolate the RNA. However, because the purification of RNA from bacterial cells is difficult, and the isolation of individual RNA species even more so, the production of RNA is usually done by transcription from cloned DNA in vitro. The most widely used

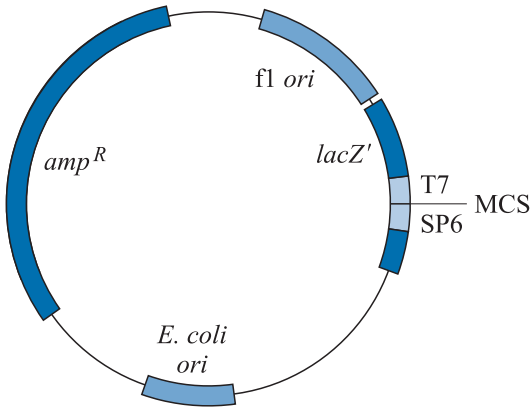


Fig 8.1 pGEMIIZf (3.2 kb). The vector contains a selectable ampicillin resistance gene (*amp*^{R}), origins of replication for double-stranded (*E. coli ori*) or single-stranded (f1 *ori*) DNA synthesis and a multiple cloning site (MCS) in a *lacZ′* gene. The MCS is flanked by T7 and SP6 promoters, allowing transcription in vitro of sequences inserted in either orientation.

systems utilize promoters from the bacteriophages T7 and SP6, but T3 promoters are also used. These phages encode novel RNA polymerases that are synthesized after infection of *E. coli*, and are very specific for phage promoters. A typical example of one of these vectors is illustrated in Figure 8.1. The vectors contain both T7 and SP6 promoters, orientated to direct transcription into the multiple cloning site that separates them. So it is always possible to transcribe a cloned sequence (regardless of the orientation in which it is cloned) by selecting the appropriate polymerase to activate the appropriate promoter. Transcription from the other promoter may be used to generate an antisense-strand RNA.

All that is needed for transcription is the incubation of the plasmid DNA in vitro with the appropriate polymerase and other substrates. Capping of the message can be achieved, if necessary, by incubation of the transcripts with the capping enzyme guanylyl transferase and GTP. If the first nucleotide to be incorporated into the transcript is a G, then it may be possible to bring about capping during transcription by the inclusion in the reaction of a cap analogue, G(5′)ppp(5′)G in excess over the normal nucleotide ppp(5′)G. The cap analogue can be incorporated only at the beginning of the transcript (because the 5′ groups, which would be needed for incorporation elsewhere, are in effect chemically blocked by each other), but it will be incorporated at the beginning of most transcripts because it is in excess over GTP. If it is important to avoid any transcription beyond the insert into the vector sequences, then the plasmid can be linearized by digestion with an appropriate restriction enzyme (cutting just beyond the site of insert integration) prior to transcription in vitro. Note that the example in Figure 8.1 has the site of insertion within the *lacZ′* section of the *lacZ* gene; therefore, as described earlier, we can verify the presence or absence of insert by the colony colour on plates containing X-gal. The plasmid also contains the filamentous phage f1 replication origin, in case we wish to generate single-stranded DNA. If cells containing the plasmid are subsequently infected with an appropriate helper phage, then

single-stranded DNA can be produced, packaged and secreted into the medium.

8.3 Synthesis of protein

8.3.1 Why synthesize protein?

It is extremely common to need to express cloned genes, so we can make the proteins they encode. We may need to prepare small quantities of radiolabelled proteins, perhaps for studies on co- or post-translational targeting or modification of proteins. Alternatively, we may want to produce larger amounts of the protein in a non-radioactive form, perhaps for determination of some of its properties in biological or biochemical assays, or for crystallization for structural studies. Or we may want to prepare an altered protein after site-directed mutagenesis to ascertain the consequences of the mutations we have generated for the catalytic properties of an enzyme.

8.3.2 The problems encountered

If we simply need to prepare small quantities of a radiolabelled protein, this is usually most easily accomplished by translation in vitro of RNA produced by transcription in vitro (see above). This is generally relatively straightforward. Occasionally, translation of small quantities of RNA is carried out in vivo, e.g. by microinjection into *Xenopus* oocytes. Larger quantities of unlabelled proteins are usually prepared by transcription and translation in vivo, and there are a number of problems to be overcome with these procedures. We need to consider the following:

(a) transcription and translation initiation
(b) fate of the transcript
(c) efficiency of translation
(d) fate of the protein after synthesis.

Methods for dealing with these considerations depend on the vectors used.

8.4 Translation in vitro

There are three commonly used methods of directing protein synthesis in vitro. They are based on a lysate of reticulocytes (immature red blood cells), a wheat-germ extract and an extract from *E. coli*. The reticulocyte lysate used comes from rabbits. Reticulocytes are prepared, lysed and the extract treated with micrococcal nuclease. This degrades the endogenous mRNA in the extract, which otherwise would produce a high background of translation products. The extract is then treated with

ethyleneglycol−O−O′−bis (2-amino-ethyl)−N, N, N′, N′−tetraacetic acid (EGTA), which chelates the calcium ions needed for the functioning of the micrococcal nuclease. If the nuclease is not inactivated, then the added mRNA will be degraded. Instead, added mRNA will be translated, and if a labelled amino acid (or acids) is included (usually radioactive methionine, although others, such as cysteine and leucine, are also frequently used) the protein produced will be labelled. The second commonly used system for directing protein synthesis in vitro is an extract of wheat germ. Basically, this is prepared by grinding wheat germ in a suitable buffer, removing cell debris, and treating with micrococcal nuclease followed by EGTA as with the reticulocyte lysate. With both these methods, transcription is usually carried out as a separate reaction (as described in section 8.2), prior to addition of the RNA to the reticulocyte or wheat-germ system. Some manufacturers produce systems in which transcription and translation systems are mixed, and the two processes are carried out together. This is also the case for the third method for translation in vitro, which uses an S-30 extract of *E. coli* cells. Cells of a suitable *E. coli* strain (usually deficient in RNase activity) are lysed, and the supernatant of a 30 000 *g* centrifuge spin is taken. This is the S-30 extract. It contains ribosomes and the other components of the protein synthesis machinery, but it does not contain the DNA, which was pelleted. Many of the ribosomes will be in the process of translating mRNA, so the extract is preincubated to allow the completion of translation that has already initiated. In prokaryotes, transcription and translation are usually coupled; that is, translation of the RNA is initiated while RNA is still being produced by transcription. Consequently, once translation of endogenous RNA has completed, it will not be reinitiated. Also, any fully transcribed RNA that is added to the system will not be translated efficiently. So, rather than transcribe RNA separately and then add it to the S-30 extract as we often do with the reticulocyte lysate or wheat-germ systems, we add DNA to the extract. Remember that the extract contains the machinery for transcription as well as translation. Any genes that have a promoter recognized by the S-30 extract (i.e. a typical *E. coli* one) can, therefore, be transcribed and translated. As with the other systems, a labelled amino acid is supplied and labelled protein is produced.

8.5 Expression in vivo

In this section we will concentrate on transcription and translation in *E. coli* cells. Similar principles apply when using other host species. We will not discuss other species in detail here, although some important considerations are detailed in the next chapter.

8.5.1 Requirements for initiation of transcription and translation

To direct initiation of transcription of a coding sequence, we need a promoter. However, it is often not enough simply to attach the coding sequence we want to express to a powerful, unregulated promoter, as many proteins are toxic to *E. coli* cells when they are over-expressed in this way. If the overexpressed protein is toxic, then no colonies will be recovered after transformation with the recombinant plasmid, because any transformed cells will be killed by the overexpression of the protein. It is possible that a few colonies might be recovered, with mutations in the cloned gene that either abolish the expression of the gene or modify the characteristics of the protein produced so that it is no longer toxic (but, in addition, no longer the protein that we are interested in). So it is usually preferable to use a powerful, but controllable, promoter.

As well as directing *transcription* initiation, we need to be able to direct *translation* initiation. In *E. coli*, this is done by a ribosome binding site. This is composed of a sequence of a few nucleotides in the mRNA, typically -GAGG-, shortly before the translation initiation codon. It is complementary to a few nucleotides at the 3′ end of the *E. coli* 16S rRNA. That sequence allows the small subunit of the ribosome to bind to the mRNA by complementary base-pairing. Translation starts at the first AUG codon in the mRNA downstream from the ribosome binding site.

So, expression of a cloned sequence in *E. coli* requires a promoter, a ribosome binding site and an initiation codon. There are a number of vectors, called expression vectors, designed to facilitate expression of cloned sequences. Examples of the strategies used are shown in Figure 8.2. Expression vectors usually contain a controllable promoter and may also provide a ribosome binding site and an initiation codon. If all three features are provided, then a hybrid protein is produced of which the N-terminal region is encoded by the vector and the rest is encoded by the sequence we inserted into the vector. The reading frame in the vector is in phase with the insert and runs directly into it. Translation will run directly from the coding region of the vector into the coding region of the insert. Proteins structured in this way are called fusion proteins. It is also possible to generate fusion proteins in which the C-terminal region is encoded by the vector. We will look at the promoters most commonly used, and then consider fusion proteins in more detail. Many expression vectors also carry powerful transcription terminators, such as that from the *E. coli rrnB* operon for rRNA, downstream from the cloning site. This is to stop high levels of transcription of other regions of the vector, such as the origin, which might interfere with the stability of the vector.

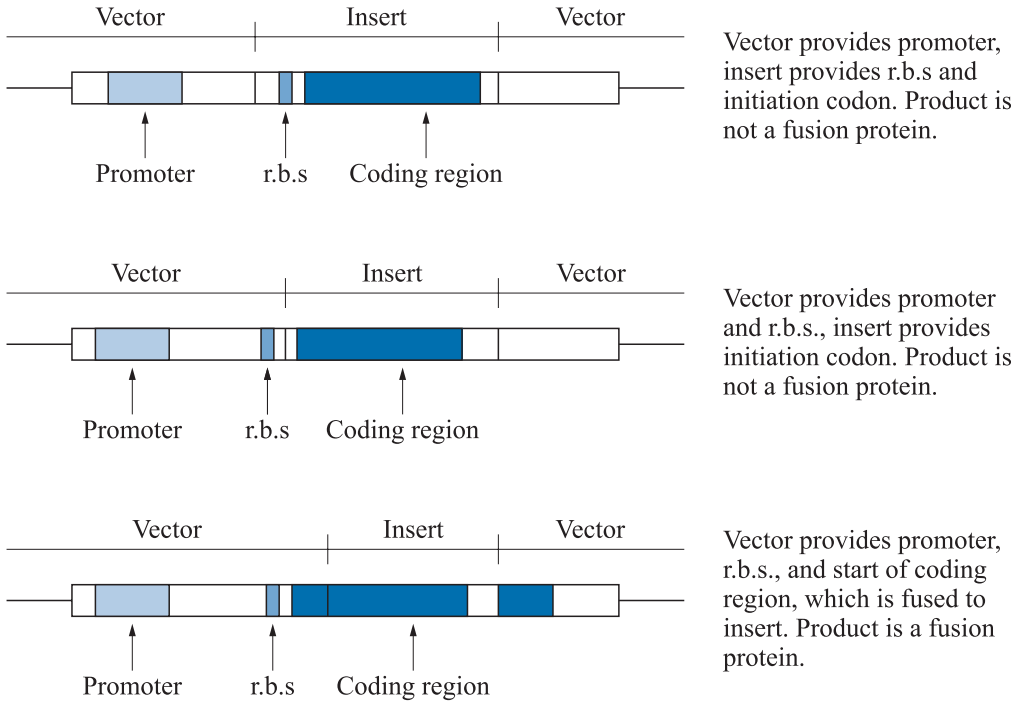


Fig 8.2 Expression of cloned sequences. r.b.s. = ribosome binding site.

8.5.2 Promoters

Several different promoters are available. Some of these occur naturally; others have been modified to improve their strength or the degree of control available. Many vectors use a two-stage control system, in which the activity of the promoter used for expression is modulated by a protein expressed from a different promoter, which is the promoter we directly control experimentally. Examples of these are given in the discussions of the lambda P_L promoter and the T7 promoter. These systems sometimes offer better control characteristics than those using a single promoter that is controlled directly. The promoters described here are available either as part of expression vectors or as cassettes that can be inserted into other vectors.

1. **lacZ promoter.** We have already come across one example of a powerful controllable promoter. This is the promoter of the *lacZ* gene, which we have already seen used in the pUC and M13 vectors, and the lambda ZAP and Bluescript series. In some vectors, a mutant form, called the *lacUV5* promoter, is used. This carries point mutations in the −10 region of the promoter, which increase its efficiency. The promoter is controlled by the LacI repressor. This is often encoded in the vector, or it can be provided by the host. In the presence of IPTG, the repressor function is inactivated and expression from the promoter occurs. Sequences inserted into the multiple cloning site in the *lacZ* gene will then be expressed. A difficulty may arise where the vector under

consideration is present in the cell at a high copy number and the repressor gene is present on an F′ plasmid, or the chromosome, at a much lower copy number. The multiple copies of the vector may be sufficient to titrate out all the repressor and allow expression in the absence of inducer. This problem of unwanted expression can be largely avoided by use of the *lacI*^{q} allele of the repressor. This carries a mutation in the promoter for the repressor itself, resulting in about 10-fold higher levels of repressor per cell than are obtained with the wild-type repressor allele. In many vectors the Lac repressor gene is carried on the vector molecule itself, which also helps to increase the amount of repressor in the cell and allows greater flexibility in the choice of host strain.

2. **Lambda P_L promoter.** The powerful leftward promoter (P_L) of bacteriophage lambda is tightly controlled by the lambda CI repressor protein acting through the O_L operator. The promoter directs transcription of phage genes during the infection process, but is widely used in expression vectors. The promoter can be controlled in a number of different ways. Perhaps the easiest way is to use a host that produces a mutant CI repressor protein (CI^{857}) that is temperature sensitive. At 30°C or lower, the repressor produced by the host is active and the promoter is kept under repression. Raising the temperature inactivates the repressor, and the promoter is activated. Another way of activating the promoter is to add nalidixic acid to the growth medium. This inhibits DNA gyrase and causes induction of the *E. coli* RecA protein, resulting in cleavage of the lambda repressor and consequent activation of the promoter. A third way of controlling the promoter is to regulate the production of the CI protein, by placing it under the control of another promoter that can be regulated. When the latter promoter is inactivated, production of CI protein stops, and expression from the lambda P_L promoter can begin. This is an example of the two-stage systems referred to in the introduction to this section.
3. **T7 promoters.** As described above, phage T7 encodes a polymerase that is very active and very specific for promoters from the same phage. A range of expression vectors has, therefore, been constructed containing phage T7 promoters to drive transcription of inserted sequences. The promoters for genes 5 and 9 of phage T7 are often used. Expression requires the production of phage T7 RNA polymerase. This is commonly supplied by a T7 RNA polymerase gene under the control of the *lacUV5* promoter. (In some systems it is under the control of a lambda P_L promoter.) Addition of IPTG activates the expression of the T7 RNA polymerase (when the *lacUV5* control system is used), and thus transcription of sequences under the control of the T7 promoter (Figure 8.3). This is another example of the two-stage systems referred to in the introduction to this section. Transcription of other genes in the host can be reduced by the addition of rifampicin, which inhibits *E. coli* RNA polymerase, but not T7 RNA polymerase.

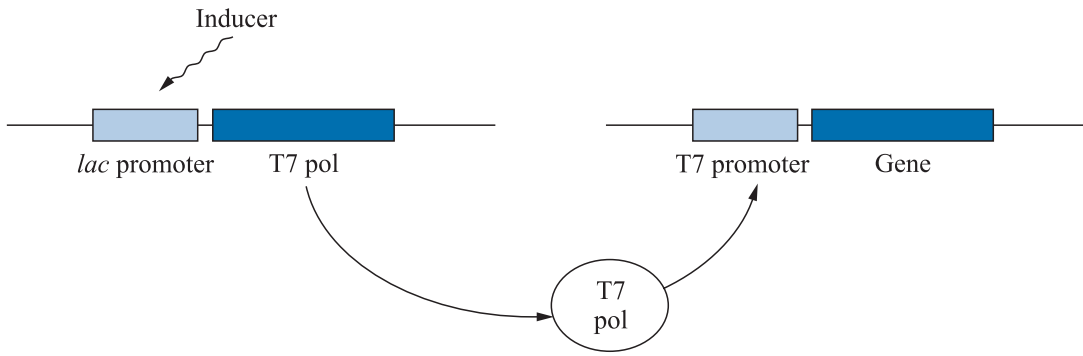


Fig 8.3 Two-stage expression systems. The target gene for expression is under the control of the T7 promoter. The T7 polymerase gene is under the control of the *lac* promoter and can be induced by IPTG. This leads to expression of the target gene.

4. **araBAD promoter.** The *araBAD* operon encodes proteins involved in arabinose metabolism. The promoter is controlled by the AraC transcriptional regulator and can be induced by addition of arabinose to the growth medium of cells containing vectors using this system. The *araC* gene is usually present on the vector.
5. **Hybrid promoters.** A number of powerful promoters have been constructed artificially as hybrids between existing promoters. One such is the *tac* promoter, which is a hybrid between the −35 region of the promoter of the *trp* operon (for tryptophan biosynthesis) and the −10 region of the *lacUV5* promoter. The *tac* promoter includes the *lac* operator and is regulated by the lac repressor, which has to be supplied by the host. Transcription is induced by addition of IPTG to the medium. Another hybrid is the *trc* promoter, which differs from the *tac* promoter in having the −10 and −35 sequences one nucleotide further apart. There is also a T7 promoter–lac operator hybrid, comprising a T7 promoter and the lac operator. Expression of this requires both T7 RNA polymerase and an inducer such as IPTG. Basal levels of expression, in the absence of the inducer, are very low indeed.

8.5.3 Fusion proteins

As explained above, expressing a sequence as an N-terminal fusion protein removes the need to provide a ribosome binding site and an initiation codon in the inserted sequence, as these are provided by the vector. Expressing a protein as a fusion may have other advantages. It may enhance stability, solubility, folding and disulphide bond formation. Antibodies to the vector-encoded part of the protein may be readily available and can be used to probe western blots and thus detect expression of the protein. Expression as a fusion to a periplasmic protein may allow the protein of interest to be targeted to the periplasmic space, which may be helpful for purification.

Affinity purification via the vector-encoded part of the protein may also be useful in purification. Several possibilities for the generation of fusion proteins are available. A short peptide region (usually of a few residues) fused to a protein of interest is often referred to as a tag.

1. **Glutathione-*S*-transferase (GST).** This enzyme catalyses the conjugation of glutathione to other molecules and has a protective function in many organisms. It is widely used as a basis for fusion proteins. The enzyme binds glutathione, allowing glutathione affinity purification matrices (such as glutathione attached to agarose) to be used very effectively in the purification of GST-fusion proteins.
2. **Maltose binding protein.** This protein is the product of the *malE* gene of *E. coli*. It is usually located in the periplasmic space, and functions in uptake of maltose and related sugars. Fusion proteins containing the leader sequence of MalE will usually be targeted to the periplasmic space, which aids in purification. Affinity purification can be carried out using a matrix containing amylose, a maltose polymer.
3. **Thioredoxin.** This protein contains two cysteine residues, which can be reversibly oxidized to cystine. The protein regulates the oxidation state of −SH groups on a range of other proteins. Fusion protein vectors are usually based on the *E. coli* gene *trxA*. Purification of fusion proteins is usually achieved by additional incorporation of a histidine tag (see below).
4. **Histidine tags.** In this example, the sequence to be expressed is fused to a sequence encoding a run of histidine residues (typically six) rather than to a naturally occurring protein. This can be either an N-terminal or a C-terminal fusion. Proteins containing histidine tags can be affinity purified using a matrix containing bound nickel ions. This is most commonly nitrilo-triacetic acid agarose (NTA-agarose). The bound nickel ions are chelated by the histidine residues of the tags (Figure 8.4).
5. **Pectate lyase.** This protein, PelB, is produced by some plant pathogenic bacteria and helps to degrade the plant cell wall. Fusion proteins produced in *E. coli* are usually secreted into the periplasm.
6. **Beta-galactosidase.** Fusion proteins incorporating beta-galactosidase can be purified using a matrix containing a covalently bound, non-cleavable, beta-galactosidase substrate analogue (such as *p*-aminophenyl-β-D-thiogalactosidyl-succinyldiamino-hexyl-Sepharose), or using a matrix containing β-galactosidase antibodies.
7. **Chitin-binding domain.** This is a 5 kDa domain from the C-terminus of a chitinase from *Bacillus circulans*. Fusion proteins carrying this domain can be purified using a column containing immobilized chitin.

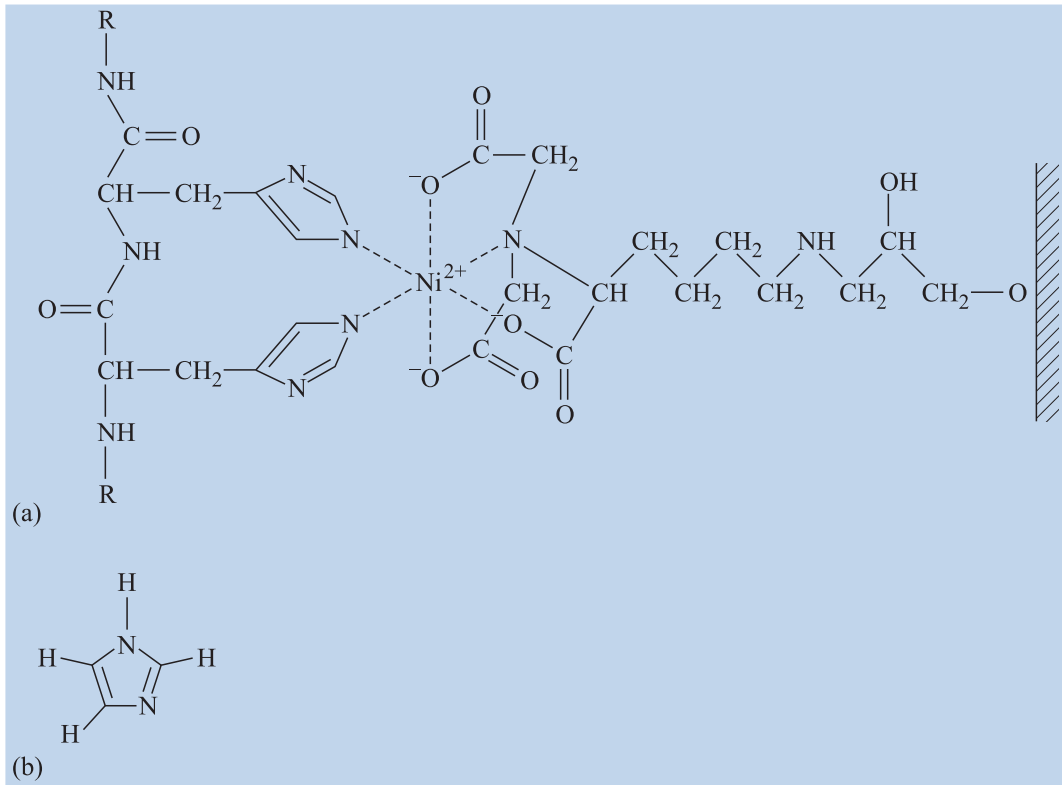


Fig 8.4 Purification of histidine-tagged proteins. (a) Histidine residues bound to nickel ions on an NTA column. (b) Imidazole, used to displace bound protein.

8. **FLAG tags.** The FLAG tag is an eight-residue peptide, DYKDDDDK. This sequence is very hydrophilic, so it is less likely to make the expressed protein insoluble. In addition, the last five residues are the recognition site for the protease enterokinase, which may be useful in removal of the tag after protein purification (see below).

Other fusion and tag systems include fusions to TrpE (involved in tryptophan biosynthesis), part of the Myc protein, the Nus protein of *E. coli*, the 'S' peptide derived from RNase A, and coloured proteins (which may facilitate monitoring of subsequent protein purification), such as haem-binding domains.

8.5.4 Purification of fusion proteins

Many fusion proteins can be conveniently purified from a cell extract by retention on a suitable affinity matrix. However, this leaves the problem of how to free the protein from the matrix. There are two common strategies for this. One strategy is simply to elute the fusion protein. This is usually done by washing the matrix with a buffer containing something that will compete for binding to the matrix or to the fusion protein. So, GST-fusion proteins can be eluted from glutathione-agarose with free glutathione, MBP-fusion proteins can be eluted from amylose matrices with free maltose, and his-tagged proteins can be eluted from NTA matrices with imidazole (see Figure 8.4), which competes with histidine for binding to the immobilized nickel. His-tagged proteins can also be eluted using

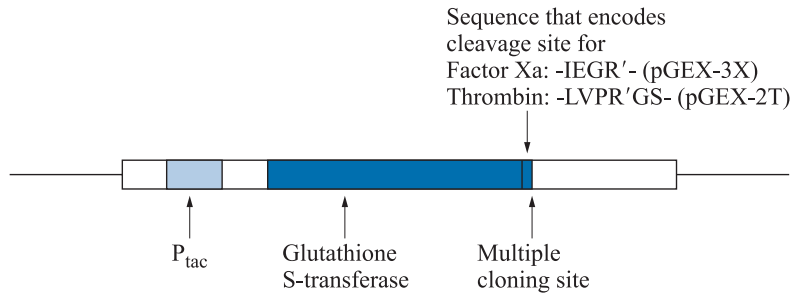


Fig 8.5 Expression regions of vectors that include protease recognition sites. The plasmids pGEX-3X and pGEX-2T (4.9 kb) are shown as examples. The vectors also have an origin of replication, an ampicillin resistance selectable marker and a *lacI*^{q} gene to supply repressor for control of the *tac* promoter (P_{tac}). The vectors differ in the cleavage site that is encoded adjacent to the multiple cloning site.

EDTA or a pH gradient to disrupt the chelation. The second strategy for releasing fusion proteins is to exploit specialized vectors that encode a protease recognition site at the junction between the fusion protein and the inserted sequence (as shown in Figure 8.5). The protease recognition sites commonly used are those for Factor Xa (-Ile-Glu/Asp-Gly-Arg-), thrombin (-Leu-Val-Pro-Arg-Gly-Ser-) or enterokinase (-Asp-Asp-Asp-Asp-Lys-). The first two proteases are involved in the blood clotting pathway; enterokinase is an intestinal enzyme that converts inactive trypsinogen to trypsin. After a fusion protein containing one of these sites has been adsorbed to a suitable matrix, the protein can then be released by treatment with the appropriate protease (Figure 8.6). It may be necessary to remove the protease subsequently, and problems can sometimes be caused by the protease cleaving the fusion protein at other sites as well. A third strategy for recovery of proteins from columns exploits the ability of certain proteins to carry out self-splicing. (This is a post-translational event, and is quite different from RNA splicing.) The splicing results in the excision of peptides called inteins and the ligation of flanking sequences (exteins). The strategy (Figure 8.7) uses an expression vector that generates a fusion protein with a modified intein sequence linking the protein of interest and the affinity tag. The intein has been modified so that it will self-cleave, on treatment with reagents (such as dithiothreitol) containing thiol groups, but will not bring about the ligation reaction. The fusion protein is purified on an appropriate affinity column and then treated to activate the intein self-cleavage. This liberates the remainder of the protein of interest.

There are also methods for chemical cleavage of the junctions within fusion proteins, e.g. with cyanogen bromide (which cleaves after methionine residues), but these are less widely used.

8.5.5 Co-expression of proteins

Sometimes it is desirable to express two or more proteins simultaneously. For example, they may form a complex that is difficult to reconstitute from the separate, independently isolated, proteins. If the proteins can be expressed in the same cell at the same time, they may assemble to form the complex, which can then be purified. There are two methods for this. One is to use separate

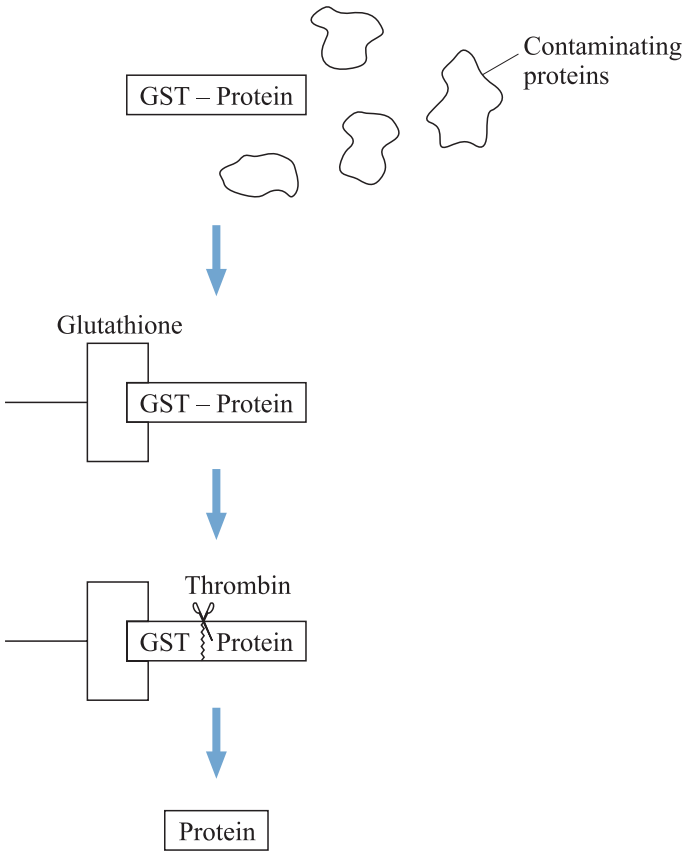


Fig 8.6 Purification and cleavage of a fusion protein. In this example, a GST fusion protein is purified by affinity chromatography on a glutathione matrix. Cleavage with an appropriate protease (thrombin in this example) releases the protein of interest.

expression vectors, with several in the cell at the same time. However, it is often difficult to maintain different, but related, plasmids in a cell stably. The second method is to use expression vectors that contain multiple independent sets of promoters and other signals in a single molecule.

8.5.6 Optimization

When expressing proteins in vivo, we usually wish to get as high a yield as possible. Very often, we also want the protein to be correctly folded and biologically active. These aims are not always compatible, so it may be necessary to find a trade-off between them. For example, we may find that if we express a protein at high levels it forms insoluble aggregates in the cell, with the protein incorrectly folded (if folded at all). A number of parameters can influence the outcome of attempts at expression, and we will look at the most important ones.

1. **Fate of the transcript.** A considerable amount of modification of the coding region of RNA molecules can take place in eukaryotes before translation. Most notable is the excision of introns, but other reactions can take place too, such as the

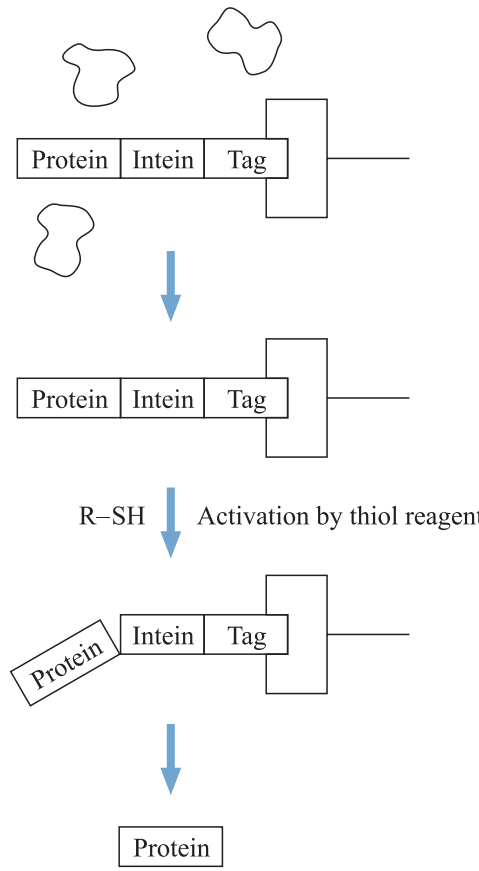


Fig 8.7 Intein-mediated cleavage of a fusion protein. The protein of interest and the affinity tag are linked by an intein domain. After affinity purification, self-cleavage of the intein is activated, liberating the protein of interest.

deletion, insertion or substitution of individual residues (editing). In general, these processes do not take place in *E. coli* (although introns are not completely absent from prokaryotes), so eukaryotic genes which normally require processing of their transcripts will not be expressed. The need for processing can be circumvented by expressing a cDNA molecule, rather than the gene itself. Assuming this cDNA has been generated from the fully processed message, there will be no need for further processing of transcripts generated from it.

2. **Efficiency of translation.** Assuming transcription and translation can be initiated efficiently, it cannot be assumed that the rest of translation will be efficient. In particular cases, strong secondary structure in a message may interfere with translation. Other factors that may influence the efficiency of translation are codon usage, codon meaning and the presence of introns (see above).

   Most amino acids can be encoded by a number of different codons, and not all codons are used equally frequently. Likewise, the tRNAs that recognize them vary in abundance. There may be differences in codon preference from one gene to another,

both within the same organism and between organisms. If the gene that is to be expressed contains many codons whose corresponding tRNA is present only at low abundance, then the overall rate of synthesis of the protein may be low, because the codons are read only with low efficiency. There are two solutions to this problem. One is to use site-directed mutagenesis to substitute the offending codons with codons that are read more efficiently. (In extreme cases, a whole gene can be synthesized to match the codon preferences of the host.) The second approach is to provide the host with a helper plasmid containing genes for the rare tRNAs. These helper plasmids are readily available commercially.

The genetic code is not universal, although deviations from it are relatively uncommon. However, if the gene we want to express comes from a system where the code is different, then expression in a system using the universal code is likely to lead to an altered product. This is perhaps most easily seen for those mitochondria (such as mammalian mitochondria) that use UGA as a tryptophan codon, rather than a stop codon. For example, expression of a mammalian mitochondrial gene containing UGA tryptophan codons in *E. coli* will lead to the formation of a truncated polypeptide, as the UGA codons are read as stop codons and protein synthesis is thereby terminated. To avoid this problem, site-directed mutagenesis can be used to change the UGA codons to UGG codons. Alternatively, one could make use of a 'chain-termination suppressor' strain of *E. coli*. Such strains contain mutated tRNA genes whose products have anticodons that are complementary to one of the termination codons. They can, therefore, 'read' a stop codon as something else, and in this particular example the UGA could be read as a tryptophan codon in an appropriate host. Because protein-coding regions frequently terminate with more than one different termination codon, the C-terminus of the protein should be unaffected.

3. **Toxicity of the protein**. Some proteins are toxic to cells when expressed at high levels. If this is so for the protein we are trying to express, then there will be a strong selective pressure for mutations in the host that inactivate expression or alter the protein itself. Therefore, it is useful to be able to keep basal levels of expression of cloned sequences as low as possible until we are ready to activate the production of our target protein. This will reduce the time available for the acquisition of undesirable mutations. Although the promoters used in expression vectors are controllable in principle, there are several addition precautions that can be taken.

   Promoters from genes involved in carbohydrate metabolism in *E. coli* are usually repressed in the presence of glucose. This is a natural physiological response to save a cell the biochemical cost of synthesizing enzymes to metabolize other sugars, when glucose is available. (The enzymes for glucose metabolism are synthesized

all the time.) So, if we are using an expression vector with the *lacZ* (or *lacZ*—T7 hybrid) or *araBAD* control systems, addition of glucose to the growth medium will help to repress basal transcription levels. When we are ready to commence expression, the cells are transferred to a medium lacking glucose.

Expression levels from T7 promoters in the absence of the T7 RNA polymerase are very low. When using a T7-based system, therefore, it may be helpful to propagate constructs in a host that lacks the T7 RNA polymerase gene. Only when we are ready to activate transcription are the constructs transferred to a host that can provide the T7 RNA polymerase. An additional way of reducing basal expression levels from T7 promoters exploits the T7 lysozyme. This phage-encoded protein has two functions. It attacks the peptidoglycan in bacterial cell walls and it also acts as an inhibitor of T7 RNA polymerase. Many of the hosts used for T7 expression vectors contain, in addition to the T7 RNA polymerase gene, a T7 lysozyme gene on a separate plasmid. The two most commonly used plasmids are called pLysE and pLysS. The former generates higher levels of the lysozyme protein in the cell than the latter. The lysozyme synthesized from these plasmids reduces the basal activity of the T7 RNA polymerase and, therefore, reduces the amount of transcription prior to induction. When synthesis of the RNA polymerase is induced, the lysozyme levels are no longer sufficient to inhibit the polymerase. A disadvantage of this system is that the reduction of T7 polymerase activity by the lysoszyme can mean that the rate of synthesis after induction is also reduced.

4. **Solubility and folding of the protein.** Many proteins expressed in *E. coli* do not remain soluble in the cell. Instead, they may form insoluble aggregates, termed inclusion bodies. This is not always a bad thing. The presence of the protein as inclusion bodies may facilitate purification, as the inclusion bodies may be easy to separate by centrifugation from the rest of the cell contents. The occurrence as inclusion bodies may also help to protect the protein against degradation, discussed in more detail below. It is possible to solubilize some proteins from purified inclusion bodies and then allow them to fold correctly. This is often achieved by solubilization using denaturing agents, followed by removal of the agents by dialysis or dilution into a buffer lacking them. The factors governing the formation of inclusion bodies are not clear. Their formation is generally favoured by higher growth temperatures and higher rates of synthesis. The protein to which the target protein is fused may also affect inclusion body formation, although if a foreign protein is synthesized as a fusion with an endogenous *E. coli* protein, then it is usually less likely to form inclusion bodies. Even if an expressed protein remains soluble, it is not guaranteed to fold correctly, or form correct disulphide bonds. If disulphide bond formation is important, then it may be

useful to direct the export of the protein to the periplasmic space. This is a more oxidizing environment than the cytosol, and thus favours disulphide bond formation.

5. **Stability.** Foreign proteins expressed in *E. coli* are often degraded. Why this degradation should occur is not always clear. It may be that if the protein concerned normally forms a complex with others, then it is unable to form a correctly folded structure in their absence. This results in partial denaturation and protease susceptibility. Formation of inclusion bodies or expression as a fusion protein often helps to maintain stability.

   A number of proteases are recognized as particularly important for the degradation of foreign proteins within *E. coli* cells. One of the major ones is the Lon protease. This protease is apparently particularly active against partially or completely denatured proteins. Inactivation of the *lon* gene can, therefore, be used to reduce the degradation of foreign proteins. This causes further difficulties, though, as the *lon* mutation can confer undesirable properties on the host cells. These include mucoidy and UV sensitivity, and sometimes reduced plasmid stability and transformation efficiency. Mucoidy is caused by overproduction of the polysaccharide capsule of the cells, making them difficult to manipulate physically. However, mucoidy can be suppressed by an additional mutation. This is often a mutation in the *galE* gene, or in the *cpsA—E* gene cluster (all of which are needed for synthesis of capsule polysaccharide). The tendency to mucoidy can also be reduced by alteration of the growth conditions, such as use of rich media and growth temperatures above 37°C. Other proteases that can lead to degradation of foreign proteins are the products of the *clpA* gene (a cytosolic protease) and the *ompT* gene (a protease in the outer membrane). Mutations in these are particularly useful in a background that is already lon^{-}. A number of proteases are under the control of the *htpR* heat-shock response gene, so mutations in this may also be helpful.

6. **Destination.** As already discussed, secretion of a protein from the cell into the periplasmic space (or even the growth medium) after synthesis may help to protect against proteolysis and will facilitate subsequent purification. Secretion may also aid in the formation of disulphide bridges. The formation of such bridges may not necessarily be an advantage, of course, depending on whether they are found in the protein in its functional state. Secretion generally (but not always) requires the presence of an N-terminal leader sequence on a protein, which is subsequently removed by a leader peptidase. In addition, for some proteins, sequences within the mature protein itself may have important effects on the efficiency of secretion. To direct the secretion of a foreign protein, therefore, we usually arrange for it to be synthesized as a fusion with the leader sequence of a secreted protein, such as MalE or PelB. Other *E. coli* proteins whose

leader sequences have been used in this way include beta-lactamase, the outer membrane proteins OmpA and OmpF, PhoA (alkaline phosphatase) and the filamentous phage gene III protein. Protein can also be exported from a cell as part of a fully formed filamentous phage, as in the phage display vectors described in Chapter 4.

7. **Modification.** Many proteins require modifications for biological activity, such as glycosylation, or the addition of cofactors, such as haem. Although prokaryotic expression systems will be able to make some of these modifications, they are unlikely to be able to carry out those that are restricted to eukaryotes. It is then necessary to use one of the eukaryotic systems discussed in Chapter 9.
8. **Other factors.** There are other factors that may be important in maximizing expression from a particular gene, particularly in industrial contexts. They include vector copy number and stability. Many of these factors are less well understood and, for most routine laboratory work using standard vectors, they are secondary considerations.

8.6 Studying gene function: reporter genes and tags

We saw in Chapter 6 how we can use reporter genes, whose activity we can readily measure, to screen libraries of cloned sequences for features such as promoters and terminators. We can also use reporter genes to study many aspects of gene function in vivo. The reporter genes most commonly used for these experiments are the fluorescent proteins (usually green, yellow, red or cyan), the luciferases, beta-galactosidase and beta-glucuronidase. The activity of all of these can be readily visualized and the activity of the different fluorescent proteins can be measured independently in the same sample. The gene for chloramphenicol acetyl transferase can also be used as a reporter, although it requires a biochemical assay, which is less convenient than direct visualization. We can also use the fusion sequences we looked at in this chapter as reporters. There are many ways reporter genes can be used. We will look at the general principles. Figure 8.8 summarizes the constructs we might make for these experiments. Many of these rely on the generation of transgenic organisms, which is discussed in Chapter 9.

8.6.1 Determining the tissue specificity of a promoter

We may want to know in which tissues of an organism the promoter of a gene of interest is active. We make a construct that places a reporter sequence under the control of that promoter and introduce the construct back into the host organism. The construct could be made either:

(a) by complete removal of the coding sequence and replacement by the reporter coding sequence, or

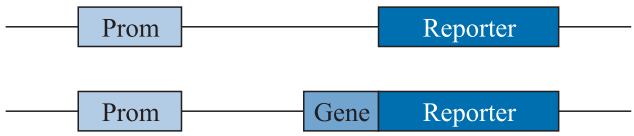


Fig 8.8 Reporter gene constructs. Both of these constructs could be used to determine promoter activity. The lower construct contains a fusion between the target gene and the reporter, and it could also be used to determine the subcellular location of the product of the target gene.

(b) so it encoded a fusion between the reporter sequence and the N-terminus of the protein whose gene we were interested in.

8.6.2 Determining the activity of altered forms of a promoter

We may have cloned a gene of interest and its associated promoter, and want to identify which regions of the promoter are important for correct expression. We make the desired changes in the promoter (Chapter 7), ligate the reporter gene to the modified promoter (either as a fusion or complete replacement, as above) and introduce the construct into the host organism. We then compare the level of expression of the reporter gene with what is obtained from an unmodified promoter.

8.6.3 Determining where a protein is located within a cell

We may want to know where a particular protein is targeted within the cell (perhaps to the mitochondrion or the chloroplast). We make a construct that contains a reporter coding sequence fused to the coding sequence of the protein of interest. Assuming (as is usually the case) the reporter sequence does not affect the targeting process, the fusion protein will be directed to its usual compartment, and we can determine this visually. It is not always easy to determine a subcellular compartment reliably when we are using fluorescence microscopy, but the ability to detect more than one kind of fluorescent protein in cells can solve this problem. We can use the green fluorescent protein as a reporter and express at the same time a red fluorescent protein whose subcellular location we know. We monitor the location of each separately. If the image of the green fluorescent protein can be superimposed accurately on the image of the red fluorescent protein, then the two proteins must be in the same place.

8.6.4 Gene trapping

We may want to identify sequences that are expressed in a particular tissue (or cell) type. A large number of transgenic organisms are created that contain a reporter gene integrated into the genome at different, random, locations. We screen the transgenic individuals for those that show the reporter gene expressed in the tissue of interest. These individuals, therefore, contain the reporter gene integrated into a sequence that is expressed in that tissue type. We can clone the sequences flanking the reporter gene using the transposon tagging approach described in Chapter 6. The use of reporter genes to identify expressed sequences in this way

is called trapping. A range of constructs is available to allow detection of enhancers, promoters and transcribed sequences, for example. These systems are then referred to as enhancer trapping, promoter trapping, and gene trapping respectively. Two examples are given in Figure 8.9. The first shows an enhancer trap system. The enhancer trap vector will not replicate independently in the host, but includes a selectable marker to allow us to select cells in which the vector has integrated into the host genome. The vector also contains a weak promoter, with a reporter gene adjacent to it. When the vector integrates near an enhancer, expression of the reporter is activated in those tissues where the enhancer is active. The second example in

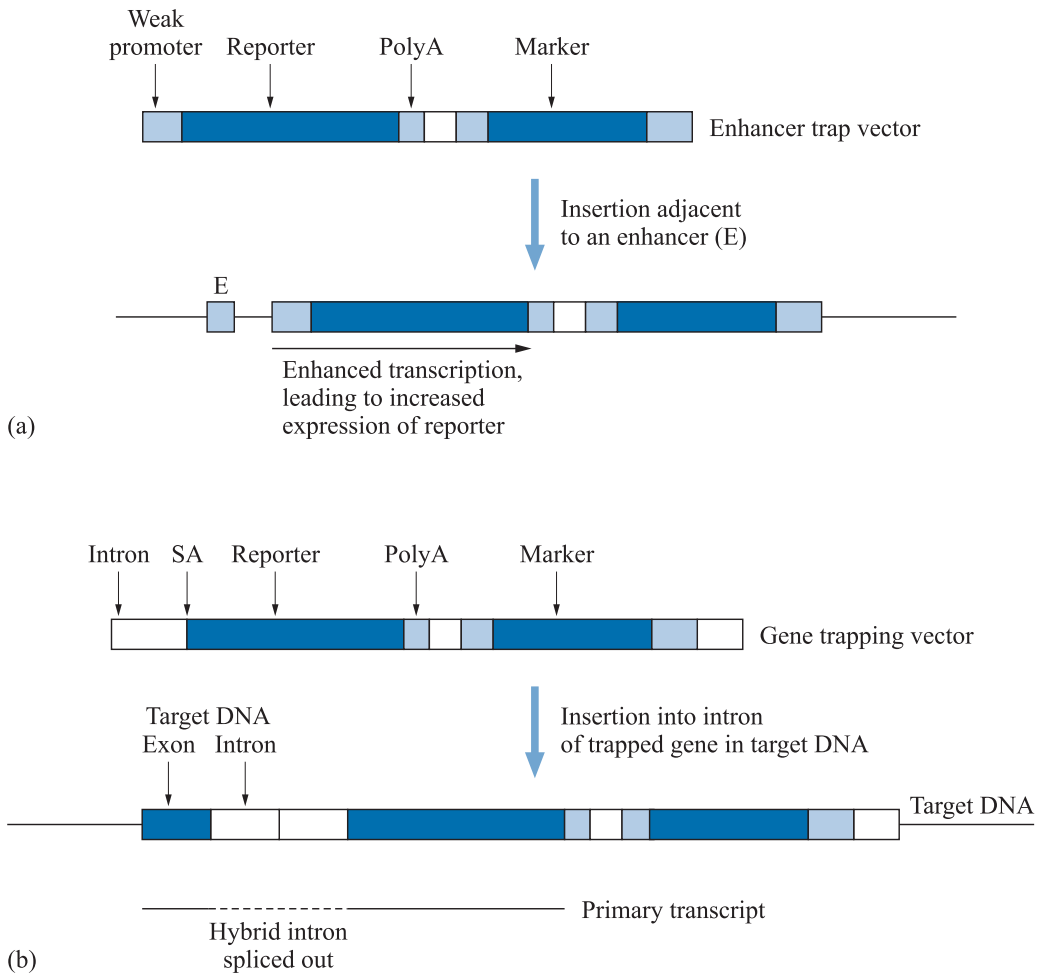


Fig 8.9 Trapping vectors. (a) An enhancer trap vector. The marker allows selection for integration of the vector. In the cases where the vector has integrated near the enhancer (E), there will be an enhanced expression of the reporter gene on the vector. (b) A gene trapping vector. The marker again allows selection for integration of the vector. In the cases where the vector has integrated within an intron (as shown), a transcript will be produced that contains a hybrid intron (dotted line), with part coming from the target DNA sequence and part from the intron in the vector. The hybrid intron will be spliced out using the splice acceptor site in the vector (SA), giving a fusion between the target DNA exon and the reporter sequence of the vector.

Figure 8.9 is a gene trap vector. Again, there is a selectable marker and a reporter gene. In this case, the reporter is promoterless and is preceded by an intron sequence. Insertion of the vector into an intron will give rise to a transcript that contains a hybrid intron, part coming from the target DNA sequence and part from the intron in the vector. The hybrid intron is spliced from the transcript using the splice acceptor site in the vector, fusing the reporter sequence to the target DNA exon.

Chapter 9

Using other organisms

9.1 Introduction

So far, we have concentrated on the techniques for manipulating the genome of *E. coli*. However, it is very likely that we will need to work with other organisms, too. We might be interested in some aspect of the biology of another species and want to study the effect of modifying it in some way. We might want to make genetically altered versions of commercially important species, such as crop plants, to improve their value. Or we might want to produce a protein in a form that requires some post-translational modification that *E. coli* is unable to accomplish. The principles that we have seen for *E. coli* apply in exactly the same way. We need to have suitable vectors, a means of getting the DNA into the organism and ways to select transformants. We may also need to take steps to increase expression. Often, we first clone DNA from one organism in *E. coli* and identify a recombinant containing a particular gene of interest. We then transfer that gene into some other host species to alter the host's properties. An organism containing a gene derived from elsewhere is said to be transgenic.

In this chapter we will look at the vectors and transformation systems available for a range of other organisms. We will look first at bacteria, and then we will consider fungi, plants (including algae) and animals. Rather than giving a detailed listing of all the methodologies for all the organisms (details of which can be found in primary research papers, reviews and laboratory manuals), we will concentrate on a representative sample of organisms to look at how the principles we have already encountered can be applied.

9.2 Bacteria

E. coli, which we have looked at in detail, belongs to the Gram-negative group of bacteria. This is just one of many bacterial

groups, all of which contain members that are important for clinical, economic, ecological or other biological reasons.

9.2.1 Vectors

The DNA used for manipulation of the species of interest could be in any of the following categories:

1. **Non-replicating DNA.** This can be linear material or circular material which, although perhaps able to replicate in *E. coli*, is not able to replicate in other hosts. It could, therefore, be stably maintained in the target species only by integration into another replicon such as the bacterial chromosome or another plasmid. If the host species has an active homologous recombination system and the incoming DNA contains sequence homologous to a site in the host genome, then insertion of the incoming DNA may take place at that site. Otherwise insertion may be at random sites, which may be disadvantageous.
2. **A plasmid unique to the species under study.** *E. coli* is, of course, not the only bacterial species to contain plasmids, and our target species may have plasmids of its own that we could use for genetic manipulation. However, it is more convenient if the plasmid used can replicate in *E. coli* as well as the target species. This would allow us to construct recombinant molecules using the well-developed systems available in *E. coli*, prior to introducing the molecules into the target species. This makes the next two categories more useful.
3. **A broad host-range plasmid.** There are many broad host-range plasmids that are able to replicate in a number of different bacterial species. They can be classified into incompatibility groups on the basis that the presence in a cell of a plasmid of one group inhibits the replication of different plasmids from the same group, as a result of interference between the processes controlling replication. There is a large number of incompatibility groups, with the five major ones being designated C, N, P-1, Q, and W. Members of the last three groups have been most widely exploited for cloning purposes, and examples are given in Table 9.1. Many have been modified by the incorporation of features such as multiple cloning sites and the *lacZ′* minigene. The plasmids differ in how easily they can be transmitted to other bacteria. Some are self-transmissible, in that they can direct their own transfer from one bacterium to another. Some of the others are unable to direct their own transfer, but can be transferred if functions are provided by another plasmid. These plasmids are said to be mobilizable. The rest cannot even be mobilized by other plasmids, and are said to be non-mobilizable. A plasmid that is transmissible may be easier to work with for that reason, but the plasmid is more likely to spread to wild populations in the event of release and, therefore, may be less desirable on safety grounds.

Table 9.1. Broad host-range plasmids of the P-1, Q, and W incompatibility groups

Plasmid	Group	Resistances[a]	Size (kb)	Transmissible?
RK2	P-1	Amp, Kan, Tet	56.4	Self
RP301	P-1	Amp, Tet	54.7	Self
pRK2501	P-1	Kan, Tet	11.1	Non-mobilizable (derivative of RK2)
pRK404	P-1	Tet (also *lacZ'*)	11.2	Mobilizable (derivative of RK2)
RSF101	Q	Str, Sul	8.9	Mobilizable
pKT254	Q	Str, Kan	12.8	Non-mobilizable
pGV1106	W	Sul, Kan	8.7	Mobilizable
pSa4	W	Cam, Str, Kan	9.4	Non-mobilizable

[a] Amp = ampicillin resistance, Cam = chloramphenicol resistance, Kan = kanamycin resistance, Tet = tetracycline resistance, Str = streptomycin resistance, Sul = sulphonamide resistance.

Table 9.2. Examples of plasmids from bacterial species and shuttle vectors derived from them

Plasmid	Source	Markers[a]	Size (kb)
pC194	*Staphylococcus aureus*	Cam	2.9
pHV33	pC194, pBR322	Cam, Amp, Tet	7.3
pIJ101	*Streptomyces lividans*	Colony morphology	8.8
pANT1200	pIJ101, pUC backbone (and other fragments)	Amp, Tsr	8.1
pUH24	*Anacystis nidulans* (cyanobacterium)		7.8
pSG111	pUH24, pBR328	Cam, Amp	12.9

[a] Amp = ampicillin resistance, Cam = chloramphenicol resistance, Tet = tetracycline resistance, Tsr = thiostrepton resistance.

4. **Shuttle vectors.** These are artificially constructed hybrids between *E. coli* plasmids and plasmids from other species. They are, therefore, able to replicate and be selected in both species. They are called shuttle vectors, as they can be shuttled from one host to another, and they are very widely used. Examples of some shuttle vectors are given in Table 9.2. In effect, they are plasmids whose host range has been artificially broadened by the introduction of an extra replication origin. Some shuttle vectors can replicate in several species. For example, the vectors based on the plasmid pC194 from *Staphylococcus aureus* can also replicate in *Bacillus* species. Typically, the initial construction and verification of a recombinant plasmid using a shuttle vector would be done using *E. coli* as a host, followed by transfer into the species of interest. There are shuttle versions of other kinds of vector, in addition to simple plasmids. These include vectors with phage elements, such as cosmids, and expression vectors. There are many markers in addition to the common ones we have already encountered. Some examples are given in

Table 9.3.	Examples of selectable markers used in prokaryotes
tsr	Resistance to thiostrepton (from *Streptomyces*)
hyg	Resistance to hygromcyin (from *Streptomyces*)
mel	Production of melanin (from *Streptomyces*)
trp	Tryptophan biosynthesis (from *Bacillus*)
sacB	Sensitivity to sucrose (from *Bacillus*)
neo	Resistance to neomycin, kanamycin (from the transposonTn5)
ble	Resistance to bleomycin (from the transposonTn5)

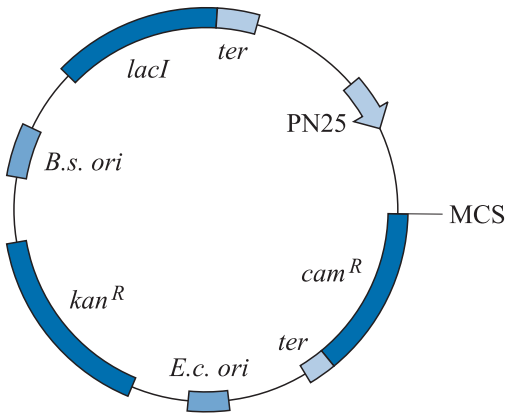


Fig 9.1 Plasmid pREP9 (7.3 kb), an example of a *Bacillus subtilis* expression vector. Sequences inserted into the multiple cloning site (MCS) are expressed from the *lac* operator–PN25 promoter (PN25). The plasmid can function as a shuttle vector, with origins for replication in *E. coli* (*E.c.*) or *B. subtilis* (*B.s.*). There are selectable markers for resistance to kanamycin (kan^R) and chloramphenicol (cam^R), and transcription terminators (*ter*).

Table 9.3. Some markers can be expressed effectively in a range of species; others may require the supply of expression signals from the species under study.

Expression vectors follow the same principles used for expression in *E. coli*. Many *E. coli* promoters will work in other Gram-negative bacteria, although powerful promoters from the species in question can also be used. Some promoters that work in *E. coli* will also work in more distantly related bacteria. For example, the P_{N25} promoter from the *E. coli* bacteriophage T5 will function in *Bacillus subtilis*, even though the latter is a Gram-positive species and *E. coli* is a Gram-negative species. If the P_{N25} promoter is fused to the *E. coli lac* operator, then the activity of the promoter can be controlled by the lac repressor in response to the addition of IPTG. The repressor is supplied by a *lacI* gene derived from *E. coli* but modified by attachment of a *B. subtilis* promoter and ribosome binding site to allow its expression in *B. subtilis*. An example of such a system, pREP9, is shown in Figure 9.1. It has a ribosome binding site and a multiple cloning site for DNA insertion, a chloramphenicol acetyltransferase reporter gene, a transcription terminator, a kanamycin resistance selectable marker, and Gram-negative and Gram-positive origins of replication (allowing use as a shuttle vector). A potential problem with expression in *B. subtilis* is

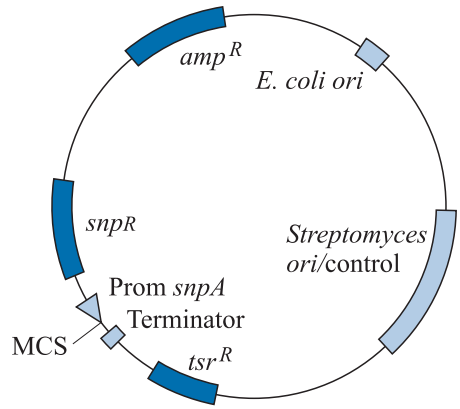


Fig 9.2 *Streptomyces* expression vector pANTI200 (8.0 kb). In addition to the ampicillin resistance gene (*amp*^{R}) and origin for selection and propagation in *E. coli*, there are an origin and an associated control sequence from *Streptomyces*, a thiostrepton resistance gene (*tsr*^{R}) and the small neutral protease promoter (Prom *snpA*) under the control of the product of *snpR*. There is a multiple cloning site (MCS) followed by a transcription terminator from the methylenomycin resistance gene.

that, during different growth stages (e.g. during sporulation), different promoters become active. Under these circumstances the expression of the promoters in a given expression vector may be considerably reduced. Homologous promoters can also be used, for expression purposes, e.g. from the *Bacillus* phage SPO-1. In *Streptomyces*, popular expression signals include promoters for *snpA* (small neutral protease), as shown in Figure 9.2, *aph* (aminoglycoside phosphotransferase), *ermE* (erythromycin resistance), *aml* (alpha-amylase) and *tipA* (for a thiostrepton-inducible protein).

Gram-positive bacteria are particularly attractive as hosts for secretion of expressed proteins into the growth medium, as they are able to do so at high rates. Secretion is usually achieved, as with *E. coli*, by expression as a fusion protein with the signal (leader) sequence of a secreted protein. Signal sequences used for *B. subtilis* include those for alpha-amylase, neutral protease, alkaline protease (subtilisin) and levansucrase. Although several of these come from different species (e.g. alpha-amylase and alkaline protease are produced by *Bacillus amyloliquefaciens*), the amino-termini of these proteins function satisfactorily in secretion by *B. subtilis*. A disadvantage with the use of *Bacillus* species as hosts for secretion is their production of high levels of endogenous proteases. Although hosts with reduced levels of protease are available, these hosts are not always satisfactory. Endogenous proteases may be less of a problem with *Streptomyces* species. Secretion signals used with these include those from tendamistat (an inhibitor of alpha-amylase, secreted by *Streptomyces tendae*) and others from proteases and protease inhibitor proteins.

9.2.2 Introduction of DNA

There are three main ways by which DNA can be introduced into recipient bacterial cells. These are the direct introduction of naked DNA into intact cells, transformation of protoplasts and

conjugation. In a few cases, and with appropriate vector systems, phage delivery may also be possible. Different methods are suitable for different species. It is important to bear in mind that many bacterial species contain restriction systems that may degrade incoming DNA. For example, these include the LT, SA and SB systems of *Salmonella*, which are analogous to the Class I *hsd* system of *E. coli*. There are Class II systems in a very wide range of bacteria, as witnessed by the very large number of different restriction enzymes available commercially. For some bacterial species there are strains lacking some of the restriction systems. Treatment of the host with UV light or high temperatures has also been used to alleviate the effects of restriction systems (although the details of why this should be effective are not always clear). In addition, artificial methylation or demethylation may be used to protect against enzymes that are active on non-methylated or methylated DNA respectively. We will now consider the different methods for the introduction of DNA into recipient cells.

1. **Introduction of naked DNA.** Many bacterial species can be induced to take up naked DNA by suitable treatments, such as the ice-cold calcium solutions or electroporation used with *E. coli*. For a species that has not previously been studied, trial and error will usually be needed to optimize the methods.

   A number of genera, including the important Gram-positive *Bacillus*, *Streptococcus* and *Streptomyces*, include species that exhibit natural competence. That is, they can take up exogenous DNA without the need for non-physiological treatment, although the ocurrence and efficiency of natural competence can vary even within species. Competence is usually regulated, and it is often determined by the excretion into the growth medium of extracellular, low-molecular-weight proteins called competence proteins. Competence then develops as cell density (and the concentration of competence proteins) in the medium increases. The requirements for transforming DNA in many of these species are complex. For example, with *Bacillus*, plasmids that are capable of direct transformation into a host generally require sequences that are also present on the host's chromosome, or sequences that are also present on another plasmid already within the cell, or at least partial internal duplication of the plasmid sequence. It is probable that entry of the molecule into the cell is associated with cutting and degradation of one strand. Regeneration of a double-stranded molecule that can be replicated may then depend on an inter- or intra-molecular recombination event. If recombination occurs between the incoming DNA and a sequence in the host chromosome or an endogenous plasmid, then the result is integration of the DNA into that chromosome or plasmid. If the incoming DNA recombines with itself in an intramolecular

reaction, then the result is recircularization. In addition, for some species the presence of a specific short nucleotide sequence on the incoming plasmid assists (and in some cases is essential for) its uptake.

2. **Transformation of protoplasts.** For several groups of bacteria, especially the Gram-positive *Bacillus* and *Streptomyces*, the constraints on the types of plasmid that can be taken up by natural transformation can be relieved by use of protoplasts (cells from which the walls have been removed by treatment with lysozyme) in the presence of a suitable osmotic buffer, to stop cell lysis, and polyethylene glycol. (Some protocols use other reagents in addition to polyethylene glycol.) Normal cells are then regenerated from protoplasts by transfer to a suitable medium. The use of protoplasts somehow bypasses the stages that cause the linearization and partial degradation of the incoming DNA. In some cases it is possible to use liposomes or, indeed, protoplasts of related species to transfer material. Disadvantages of protoplast transformation include the difficulty of regenerating normal cells from protoplasts for some species and the fact that the rich media used for this may preclude the selection of markers (especially nutritional ones).
3. **Conjugation**. In conjugation, two cells become physically linked by a proteinaceous tube (or pilus) and DNA is passed from one cell to the other. This is how the F (for 'fertility') plasmid, which forms the basis of a lot of *E. coli* genetics, is transferred from one cell to another. The F factor encodes all the functions necessary for its own transfer. However, other, smaller, plasmids exist that are more suitable for artificial transfer between species. These are the broad host-range plasmids referred to above. The plasmid to be transferred (sometimes called the cargo plasmid) is often unable to direct its own transfer. Instead, it relies on a conjugal plasmid that is able to do so, and sometimes one or more helper plasmids that encode functions such as methylases that will protect the cargo from degradation in its future host. The helper plasmid may also carry functions needed for the transfer of the cargo. Typically, the transfer is achieved in a triparental mating (Figure 9.3), of the following three conjugants: *E. coli* carrying the conjugal plasmid; *E. coli* carrying the cargo and any helper plasmids; and the recipient bacterial species to be manipulated. Conjugation between the two strains of *E. coli* places all the plasmids necessary in one *E. coli* cell, and conjugation with the recipient bacterium transfers the cargo into the recipient species. There are variations on this approach. If all the plasmids can be assembled in one *E. coli* strain first, then a biparental mating is sufficient to introduce the plasmid of interest into the recipient strain. Although *E. coli* can conjugate with a wide range of bacteria, including some Gram-positive bacteria, other bacterial species can be used as a donor. If the same species can be used as donor and recipient, then

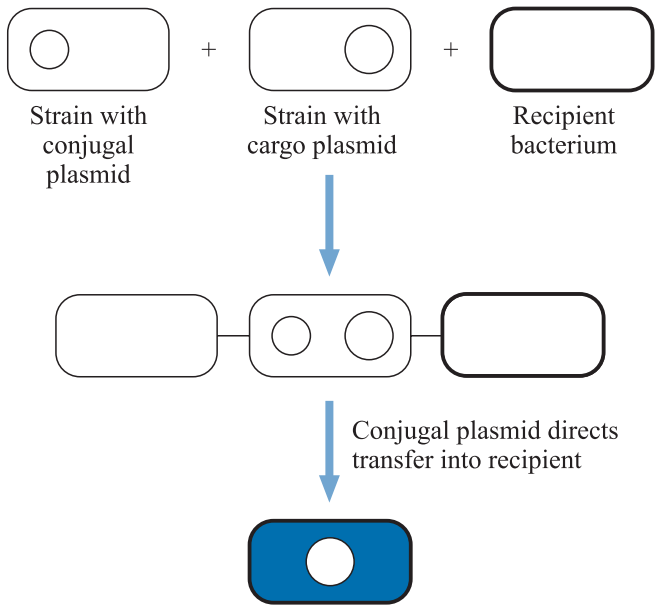


Fig 9.3 Triparental mating to mobilize a cargo plasmid into a recipient.

the plasmid transfer may be more efficient. Conjugation is often useful where direct transformation and protoplast transformation have failed.

9.2.3 Hosts

The same considerations for host design in *E. coli* apply to other bacteria. *B. subtilis* probably offers the best range of host strains. In this species, in addition to host nutritional markers, such as *trpC2*, *leuB6* and *metB10* (blocking tryptophan, leucine and methionine biosynthesis respectively), there are mutations available (such as *recE4)* that lead to reduction in recombination and plasmid instability. There are also mutations reducing endogenous restriction enzyme activity and protease activity (which may be useful when expressing cloned genes). Useful mutant strains are available for some other species, although genes such as *recA* that can be inactivated in some species may be essential (or very nearly essential, so that inactivation leads to very low viability) in others.

9.3 *Saccharomyces cerevisiae*

The fungi are an extremely diverse group of organisms. Perhaps the best studied is the ascomycete fungus *Saccharomyces cerevisiae* (also known as bakers' yeast, or simply yeast), so we will concentrate on that. However, vectors and transformation systems have been

Table 9.4. Selectable markers for use in *Saccharomyces cerevisiae*

Marker	Function
URA3	Orotidine-5′-phosphate decarboxylase (in *de novo* synthesis of pyrimidines)
LEU2	β-Isopropylmalate dehydrogenase (in leucine biosynthesis)
HIS3	Imidazole glycerol phosphate dehydratase (in histidine biosynthesis)
ARG4	Argininosuccinate lyase (in arginine biosynthesis)
TRP1	*N*-(5′-Phosphoribosyl) anthranilate isomerase (in tryptophan biosynthesis)
TUN^R	UDP-*N*-acetylglucosamine-1-P transferase (in glycosylation; confers tunicamycin resistance)
TCM1	Ribosomal protein (confers resistance to trichodermin)
CYH2	Ribosomal protein (confers resistance to cycloheximide)
ADE2	Phosphoribosyl amino-imidazole carboxylase (in *de novo* biosynthesis of purines)
CAN1	Permease (confers sensitivity to canavanine)
SUP4	Tyrosine-tRNA (ochre chain termination suppressor)

developed for a number of other fungi, including species that are of particular interest as pathogens for plants and animals, so we will look briefly at some of those as well. We will look at the markers available, the vectors containing them and the ways of getting DNA into host cells.

9.3.1 Markers

One of the most commonly used selectable markers is the *URA3* gene. This encodes orotidine-5′-phosphate decarboxylase, an enzyme of uracil biosynthesis. Acquisition of a functional *URA3* gene by cells that were previously mutant at this locus allows them to grow in the absence of exogenous uracil. Examples of other frequently used selectable markers, including nutritional markers that can be selected in a similar way, are given in Table 9.4. Some of the markers (such as *CAN1*, *SUP4* and *URA3*) can be counterselected, i.e. we can select for their absence. The *CAN1* gene encodes a permease that causes uptake of the toxic arginine analogue canavanine and consequent cell death. So, growth of cells in the presence of canavanine imposes selection for those lacking a wild-type *CAN1* gene. This also forms the basis of selection for *lack* of *SUP4*, using a host with an ochre chain termination mutation in the *CAN1* gene. The *can1* mutant host cells are resistant to canavanine, as they are unable to import it. However, the chain termination suppressor *SUP4* causes the production of functional permease in the *can1* background (because it suppresses the *can1* mutation) and, therefore, causes the death of *can1* cells in the presence of canavanine. Loss of *SUP4* abolishes canavanine uptake and causes canavanine resistance, which can readily be selected. Cells containing *SUP4* can also be identified (without killing them) by the use of hosts with an ochre chain termination mutation in the *ADE2* gene for phosphoribosyl amino-imidazole carboxylase, a component of the

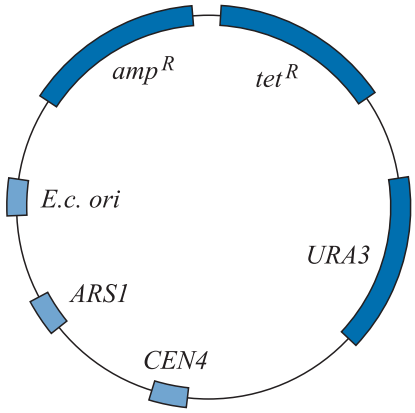


Fig 9.4 Yeast centromeric plasid YCp50 (7.9 kb). The presence of a suitable origin (*E. c. ori*) and selectable markers (amp^R and tet^R) allows it to be used as a shuttle vector in *E. coli*. The vector contains a yeast selectable marker (*URA3*), a centromeric sequence (*CEN4*) and an autonomously replicating sequence (*ARS1*).

purine biosynthesis pathway. The mutation causes the cells to accumulate a red pigment, whereas colonies are the usual colour if the ochre mutation is suppressed by *SUP4*. Cells that have lost *SUP4*, therefore, acquire the red pigment and are visibly distinguishable from the others. Cells mutant in *URA3* can be selected using 5-fluoro orotic acid (5-FOA), which is turned into toxic products by the wild-type protein.

9.3.2 Plasmid vectors

Integration of non-replicating incoming circular or linear DNA into the yeast genome by homologous recombination can be used to bring about stable genetic modification. The need for integration will clearly make recovery of the incoming DNA much more difficult than if a replicating vector is used. Integration is the basis of gene disruption, widely used for inactivating genes in yeast (see Chapter 7).

The replicating plasmid vectors used commonly in yeast can be divided into two main categories: yeast centromeric plasmid (YCp) and yeast episomal plasmid (YEp) vectors. An example of a YCp plasmid is given in Figure 9.4. It contains prokaryotic ampicillin and tetracycline resistance genes and an origin of replication for *E. coli*, all in a region of the plasmid derived from pBR322. These will allow it to function as a shuttle vector, and sequences can be cloned into the ampicillin or tetracycline resistance genes. Many of the yeast vectors are based on pBR322 in this way, but other plasmids have also been used including the pUC and Bluescript species. There are also a selectable marker and an origin of replication for yeast, *ARS1*. (An ARS, or autonomously replicating sequence, confers on a plasmid the ability to replicate without integration into another replicon. There are several available from yeast.) The distinguishing feature of YCp plasmids is the presence of a chromosomal centromeric sequence, *CEN4* in this example. The presence of a centromeric sequence allows the plasmid to be partitioned by the spindle apparatus of the cell, in the same way that the endogenous chromosomes are partitioned. Although this also results in a large reduction of the copy number per cell, it confers a much greater stability during cell division, allowing

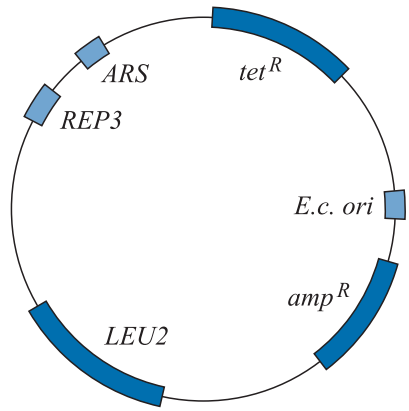


Fig 9.5 Yeast episomal plasmid YEpl3 (10.8 kb). The presence of a suitable origin (*E. c. ori*) and selectable markers (amp^R and tet^R) allows it to be used as a shuttle vector in *E. coli*. The vector contains a yeast selectable marker (*LEU2*), a partitioning sequence (*REP3*), and an autonomously replicating sequence (*ARS*) derived from the 2 μ plasmid.

the plasmid to be maintained in the absence of selection. Plasmids carrying an *ARS* but no *CEN* sequence are sometimes called YRp (yeast replicating plasmid) vectors. Although they have higher copy numbers than YCp vectors, their instability is a major drawback.

An example of a YEp vector is shown in Figure 9.5. Most of the features are similar to those of YCp plasmids, except for the absence of the centromeric sequence and its replacement with a different origin of replication. This comes from a naturally occurring episome (plasmid) from yeast, called the 2 μ (micron) circle. As well as the origin of replication, there is a naturally occurring *cis*-acting sequence, *REP3*, which is the site of action for proteins that help in the partitioning of the plasmid at cell division. This gives the YEp vectors a mitotic stability similar to that of the YCp vectors, but with a somewhat higher copy number. It may be possible to increase copy number further by use of a selectable marker (e.g. some alleles of *LEU2*) that complements only poorly. This gives a selective advantage to cells with a higher plasmid copy number.

There are more sophisticated versions of these classes of vector, with further modifications. These include the presence of filamentous phage origins for generation of single-stranded DNA and the *lacZ′* system for detecting the presence of inserts when the vectors are used in *E. coli*.

9.3.3 YACs

YACs are sophisticated cloning vectors that can be used for propagating large stretches (in the megabase size range) of DNA. This is very useful, because it reduces the number of recombinants needed to cover the entire genome of an organism. The aim is to construct an artificial chromosome that can be maintained in yeast, but which is mostly composed of foreign DNA. An example is shown in Figure 9.6 together with a representative cloning strategy. The vector contains *HIS3*, *URA3*, *TRP1* and *SUP4*, described above. *CEN4* is a centromeric sequence and *TEL* is a telomeric sequence derived from the ends of ribosomal RNA-encoding molecules from the macronucleus of the protozoan *Tetrahymena*. In addition, there

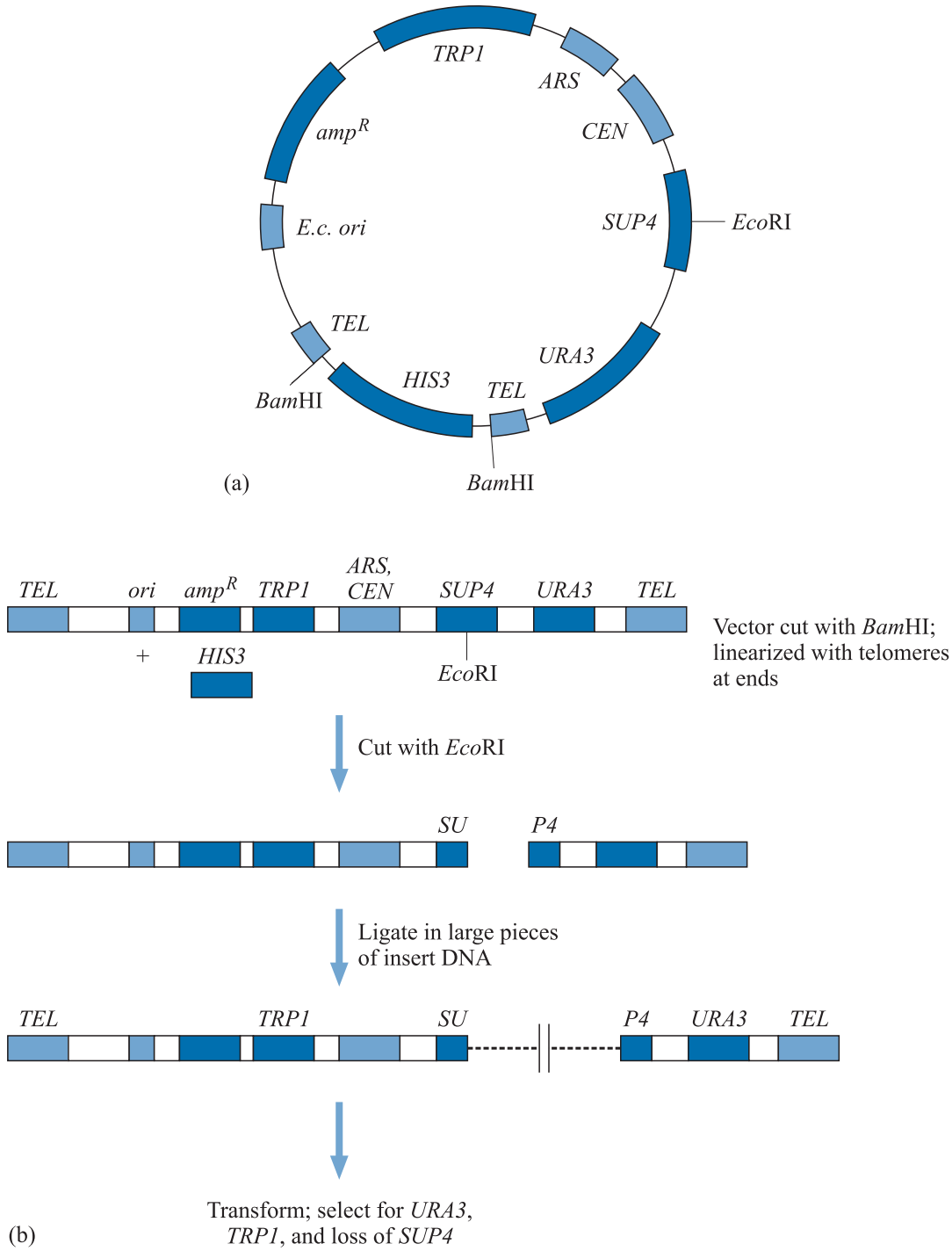


Fig 9.6 YAC cloning: **(a)** shows a representative YAC, pYAC4; **(b)** a typical cloning strategy. Note the telomeres (*TEL*) that constitute the ends of the linear recombinant molecule, as shown in panel **(b)**. Other sequences are as described in the text. The inserted DNA is indicated by the broken dashed line.

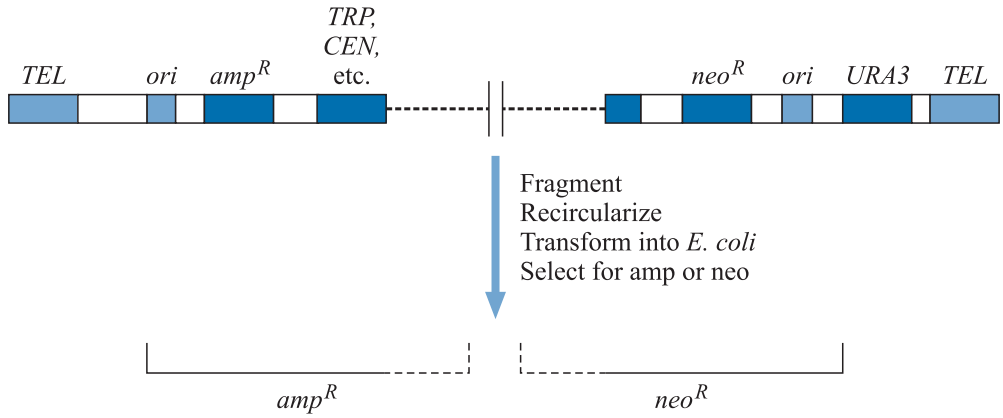


Fig 9.7 YAC cloning and recovery in *E. coli*. Note the presence of a prokaryotic origin of replication (*ori*) and drug resistance gene (amp^R or neo^R, for resistance to ampicillin and neomycin) in each arm, allowing recovery of each arm and adjacent inserted sequences.

are a yeast origin of replication, a prokaryotic origin of replication, and an ampicillin resistance gene. The latter two features allow YAC DNA to be prepared from *E. coli* prior to the cloning. The whole molecule is just over 10 kb.

The YAC DNA is digested, in this case with *Bam*HI and *Eco*RI; this generates three fragments, two of which contain *TEL* sequences and which are the ones we will need. (The third has the *HIS3* gene.) Subsequent religation of the fragments can be prevented by treatment with phosphatase. The DNA is then mixed with the insert material (which has been prepared in a way that leaves *Eco*RI ends) and ligated and the ligation products are introduced into yeast. There, they can function as *bona fide* chromosomes, possessing a centromere, an origin of replication and telomeres. Transformed cells can be selected by the presence of the *URA3* and *TRP1* genes. The presence of both of these markers indicates that both arms of the YAC have been acquired. Screening for the inactivation of *SUP4* allows putative recombinants to be identified.

There are several modifications to this basic YAC design. One is the incorporation of an extra prokaryotic origin and prokaryotic selectable marker, so that there is one of each on each arm. This assists in the recovery of at least parts of individual recombinant YACs using *E. coli*. The YAC DNA is isolated from yeast, digested, self-ligated and used to transform *E. coli*. Cells containing either the left arm or the right arm can then be selected (Figure 9.7). Bacteriophage promoters (e.g. from T7) can also been included to allow the synthesis of RNA transcripts in vitro.

9.3.4 Expression systems

A number of constitutive promoters are available, e.g. from the *GPD* gene for glyceraldehyde-3-phosphate dehydrogenase, *TEF* for translation elongation factor 1α, and *CYC1* for a component of cytochrome *c*

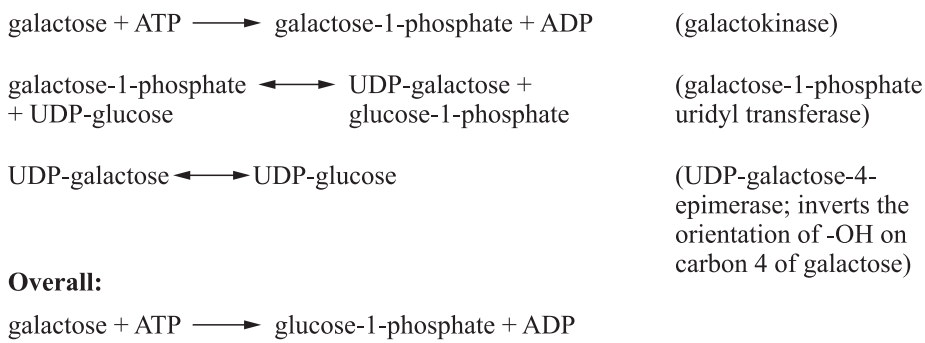


Fig 9.8 Metabolism of galactose. The glucose-1-phosphate can be converted to glucose-6-phosphate through glucose-1,6-bis-phosphate and then metabolized by glycolysis or other standard routes.

Table 9.5. Regulatable promoters for use in *Saccharomyces cerevisiae*

Promoter	Function and control
GAL1	Galoctokinase; upregulated in presence of galactose, repressed in presence of glucose
GAL10	UDP-galactose-4-epimerase; upregulated in presence of galactose, repressed in presence of glucose
PGK	3-Phosphoglycerate kinase; upregulated in presence of glucose, repressed on non-fermentable carbon sources, such as acetate
ADH1	Alcohol dehydrogenase; upregulated in presence of glucose, repressed on non-fermentable carbon sources, such as acetate
PHO5	Secreted acid phosphatase; upregulated in medium deficient in inorganic phosphate (to assist scavenging), repressed in presence of phosphate
MET25	O-Acetyl homoserine sulphydrylase (enzyme of methionine biosynthesis); repressed in presence of methionine
CUP1	Copper detoxification; upregulated when copper present in growth medium
CTR1, CTR3	Copper transport; upregulated when copper absent from medium

oxidase. Some of the most widely used controllable promoters come from genes for galactose metabolism (Figure 9.8), *GAL1* and *GAL10*, encoding galactokinase and UDP-galactose-4-epimerase respectively. When wild-type control proteins are present, transcription of *GAL1* and *GAL10* is increased greatly in the presence of galactose. In the presence of glucose, transcription is tightly repressed. There are two deletion variants of the *GAL1* promoter, designated *GALL* and *GALS*, that are less strong and, therefore, less likely to lead to problems caused by toxicity of foreign proteins. All these promoters, together with their upstream controlling sequences, can be used either in expression vectors or as cassettes for insertion into other constructs. These and other promoters that can be tightly controlled and are used for expression purposes are summarized in Table 9.5. All these promoters can be used to direct the synthesis of either unmodified proteins or fusion proteins; the latter can incorporate

either the corresponding products of *GAL1*, *GAL10* and so forth, or other sequences, such as LacZ or TrpE from *E. coli*. An advantage with using yeast rather than prokaryotic cells for expression of eukaryotic proteins is that post-translational modifications of the proteins (such as glycosylation, phosphorylation, and acylation) are unlikely to take place in prokaryotic systems but they may do so in yeast. Even in yeast, these modifications may not take place reliably, and it may be necessary to use expression in insect or mammalian cells (see below).

9.3.5 Secretion systems

Secretion of overexpressed proteins by fusion to a secretion signal may be useful because it assists purification and allows the proteins produced to be only minimally exposed to most of the yeast intracellular proteases. In addition, it may help to bring about correct glycosylation and may also enhance protein folding. A number of secretion signals, from both yeast and other organisms, can be used. Those for the mating pheromone alpha-peptide and for invertase (product of the *SUC2* gene) are often used. The mating pheromone is a 13-residue peptide, secreted from cells of alpha mating type, that acts on a receptor on cells of mating type a. Invertase is secreted in order to bring about the breakdown of sucrose to glucose and fructose.

9.3.6 Transformation systems

Three ways are commonly used for introducing exogenous DNA into *S. cerevisiae*. One of the simplest and most widely used is the treatment of cells with lithium acetate and polyethylene glycol and a heat shock, together with the transforming DNA and additional non-specific denatured 'carrier' DNA (e.g. from calf thymus). This method is relatively fast and straightforward and can give an efficiency of transfer of up to 10^5 to 10^6 colonies per microgram of plasmid with suitable strains.

A second approach to introducing exogenous DNA into *S. cerevisiae* is the transformation of protoplasts (also called spheroplasts), which are obtained by digestion of the cell wall with appropriate enzymes, in the presence of a suitable osmotic buffer, such as sorbitol solution, to prevent lysis. The protoplasts are exposed to the transforming DNA, in the presence of polyethylene glycol and calcium chloride. The next step is to regenerate intact cells from protoplasts. This is achieved by immobilizing the protoplasts in a suitable agar that is osmotically buffered. Surrounding the protoplasts with a solid medium allows the retention of cell wall material around the cell during the earliest and most sensitive stages of cell wall regeneration. Protoplast transformation is more time consuming than transformation with lithium acetate, but with some strains it can give a higher transformation efficiency. Therefore, it is useful

with strains that otherwise have a low transformation efficiency, or when the amount of DNA available is low.

A third method for introduction of DNA into *S. cerevisiae* is electroporation. It is reported that this method can give even higher efficiency than protoplast transformation, although the actual figure seems to vary quite widely and probably depends on the strain used. A disadvantage is that the number of transformants obtained reaches a plateau with relatively low amounts of DNA. As with other electroporation systems, the cells are simply subjected to an electric field pulse. There is no need to make protoplasts for this treatment, although an osmotic buffer such as sorbitol is often included. Treatment with lithium acetate and dithiothreitol is also sometimes included to increase transformation efficiency.

Other less widely used methods for transformation of *S. cerevisiae* include particle bombardment and cell damage by agitation with glass beads. Transmission of DNA from *E. coli* by conjugation has also been reported. The *E. coli* cells must contain a plasmid that can replicate in *E. coli* and in yeast (i.e. a shuttle vector) and another plasmid to provide mobilization functions.

9.4 Other fungi

9.4.1 *Schizosaccharomyces pombe*

Another unicellular fungus that is frequently studied is the fission yeast *Schizosaccharomyces pombe*. It is particularly widely used in work on the cell cycle and eukaryotic chromosome biology. As its name implies, *Schizosaccharomyces pombe* reproduces by fission, in contrast to *Saccharomyces cerevisiae*, which reproduces by budding. Many of the early vectors for use with *Schizosaccharomyces pombe* contained the origin of replication from the 2μ plasmid of *Saccharomyces cerevisiae*. However, this does not always allow stable maintenance in *Schizosaccharomyces pombe*, and a more common strategy now is to use plasmids with a *Schizosaccharomyces pombe* origin of replication, *ars1*. However, because the *Schizosaccharomyces pombe* centromere is very large, it is not possible to include it in such plasmids. Consequently, the copy number of these plasmids can vary widely and the plasmids may be lost altogether. Therefore, it may be better to direct the integration of incoming DNA, which can be achieved either by homologous or by non-homologous recombination.

A number of selectable markers are available. One of the most widely used is *ura4*$^{+}$, which encodes orotidine-5′-phosphate decarboxylase (the enzyme encoded by *URA3* of *Saccharomyces cerevisiae*). This can be selected in a *ura4*$^{-}$ background in the usual way. (The *URA3* gene of *Saccharomyces cerevisiae* is sometimes used, but is less efficient.) Absence of *ura4*$^{+}$can be selected using 5-fluoro-orotic acid, which is converted to a toxic product by the

enzyme. The *leu1*$^+$ marker, encoding β-isopropylmalate dehydrogenase, is also popular. As with *ura4*$^+$, the *Saccharomyces cerevisiae* equivalent (*LEU2*) is also used, but is less efficient. The *sup3–5* gene of *Schizosaccharomyces pombe* is often used, as a suppressor of a mutation in the *ade6* locus (equivalent to *ADE2* of *Saccharomyces cerevisiae*). Markers based on genes for histidine biosynthesis (*his3*$^+$ and *his7*$^+$), arginine biosynthesis (*arg3*$^+$) and antibiotic (G418 or hygromycin) resistance are also available. There are counterselection systems for loss of the *Saccharomyces cerevisiae CAN1* gene and a thymidine kinase gene from herpes simplex virus, as well as the *ura4*$^+$gene. The replacement of one plasmid by another by introduction of the second plasmid into a strain containing the first, followed by removal of the first by counterselection is sometimes called a plasmid shuffle.

Expression vectors are available with a range of constitutive or inducible promoters. Constitutive promoters include that from the *adh1*$^+$ alcohol dehydrogenase gene, as well as those for the small-T antigen transcript of SV40 and the 35S transcript of the plant virus CaMV. (The last of these has been modified to contain a tetracycline control system derived from a bacterial transposon, so that expression is induced in response to added tetracycline.) One of the most widely used inducible promoters is from the *nmt1*$^+$ gene. This gene presumably encodes a polypeptide involved in thiamine bioysnthesis, as there is *n*o *m*essage in *t*hiamine. There are various forms of this promoter, which have been artificially modified to provide a range of promoter strengths. The promoter from the *fbp1*$^+$ gene for the gluconeogenic enzyme fructose 1,6-bis-phosphatase is also widely used. The promoter is tightly repressed in the presence of glucose, but activated when glucose is absent and glycerol is present as a carbon source instead (conditions when the gluconeogenesis pathway is active). Other regulated promoters include those for invertase (induced by growth in the presence of sucrose and absence of glucose) and a copper transporter (induced by growth in the absence of copper).

Early work on *Schizosaccharomyces pombe* used protoplast transformation for the introduction of DNA. However, this has now largely been replaced by electroporation and transformation using lithium acetate. This is probably because protoplast transformation is more time consuming, and it is also reported to favour non-homologous integration when using non-replicating sequences.

9.4.2 Other yeasts

Other yeasts may be useful in having more efficient secretion systems than *Saccharomyces* and *Schizosaccharomyces*. These include *Kluyveromyces lactis*, *Hansenula polymorpha* and *Pichia pastoris*. There are integration vectors available for *Kluyveromyces lactis*, and also vectors for extrachromosomal replication using chromosomal ARSs or origins from naturally occurring *Kluyveromyces* plasmids.

Secretion can be directed in *Kluyveromyces* by fusion to suitable sequences from a secreted toxin that kills other cells. The yeasts *Hansenula polymorpha* and *Pichia pastoris* are methylotrophic. That is, they can grow on methanol as a carbon source, and this induces the expression of genes for enzymes such as methanol oxidase (also known as alcohol oxidase) and formate dehydrogenase. This induction has been used as the basis for expression vectors. Proteins can be secreted at high levels using a range of signal sequences, including the signal sequence from the alpha mating factor of *Saccharomyces cerevisiae*, and there is generally little proteolysis of secreted material. The methanol oxidation pathway depends on the peroxisome, and these yeasts have also been exploited for the study of targeting to the peroxisome.

9.4.3 Filamentous fungi

There is a lot of interest in the transformation of other fungi (and other organisms loosely classified as fungi, such as oomycetes). Filamentous fungi are of particular importance, as they include a wide range of plant and animal pathogens and biological control agents. Techniques for the genetic manipulation of the model filamentous fungi *Neurospora crassa* and *Aspergillus nidulans* were developed not long after those for *Saccharomyces cerevisiae*. Methods are now available for a wide range of such organisms, including *Acremonium chrysogenum* (used for cephalosporin production), *Aspergillus oryzae* (used in a range of fermentations), *Fusarium oxysporum* (a pathogen with a broad host range), *Magnaporthe grisea* (causative agent of rice blast disease), *Penicillium chrysogenum* (used for penicillin production), *Trichoderma viride* (a wood-digesting fungus also used for biological control), *Ustilago maydis* (causative agent of smut disease on maize) and the oomycete genus *Phytophthora* (including the pathogen responsible for the Irish potato famine in the mid-19th century).

A wide range of selectable markers is available. They are summarized in Table 9.6. Some of these require isolation of an appropriate mutant strain in which to select the marker. Strains deficient in orotidine-5′-decarboxylase can be isolated by virtue of their resistance to 5-fluoro-orotic acid, and strains deficient in nitrate reductase can be isolated by their resistance to chlorate (which generates a toxic product under the action of nitrate reductase). Many markers, including their promoters, can be used across a wide range of species. So, for example, the *argB*$^+$ gene from *Aspergillus nidulans* was used in early work on *Magnaporthe grisea*, and the *niaD*$^+$ gene of *Aspergillus nidulans* was used in a *niaD*$^-$ strain of *Fusarium oxysporum*. Many promoters (such as heat shock promoters) and origins of replication function effectively in a wide range of fungi. The origin of replication of the *Saccharomyces cerevisiae* 2 μ plasmid can also be used in a number of filamentous species. In some cases, stable maintenance of extrachromosomal molecules is

Table 9.6. Selectable markers commonly used in filamentous fungi

Marker	Source	Product	Phenotype
argB	*Aspergillus nidulans*	Ornithine trancarbamoylase	Ability to synthesize arginine
hph	*Escherichia coli*	Hygromycin phophotransferase	Hygromycin resistance (phosphorylates it)
ble	*Streptoalloteichus hindustanus*	Bleomycin-binding protein	Bleomycin resistance (binds to it and inhibits its DNA-damaging activity)
pyrG	*Aspergillus nidulans*	Orotidine-5′-decarboxylase	Ability to synthesize uracil (= *URA3* of *Saccharomyces cerevisiae*, and *pyr-4* of *Neurospora crassa*)
niaD	*Aspergillus nidulans*	Nitrate reductase	Ability to grow with nitrate as sole nitrogen source
benA	*Aspergillus nidulans*	Beta-tubulin	Resistance to fungicide benomyl
amdS	*Aspergillus nidulans*	Acetamidase	Growth on acetamide as sole carbon or nitrogen source

difficult, because of the filamentous nature of the fungi (so that cells that have lost a selectable marker can be maintained by those retaining it), so integration is more convenient. DNA may be integrated into host chromosomes by homologous or non-homologous recombination.

There are a number of methods for the introduction of DNA. In early work, transformation was carried out with protoplasts made from spores or hyphae by digestion of the cell wall with suitable enzymes. DNA is added in the presence of calcium ions and polyethylene glycol. The latter apparently causes protoplast fusion, and DNA is taken up at the same time. The protoplasts are then regenerated. Another method is to expose germinating spores to lithium acetate and polyethylene glycol together with the transforming DNA. This circumvents the need for making and regenerating protoplasts. Electroporation is also commonly used, either with protoplasts or with other cell types. Where cells have thick walls and protoplast production is difficult, biolistic transformation is often useful. In many cases, if integration of a marker is selected, it is found that other, non-selected, DNA species that were applied at the same time are also integrated. This is called co-transformation. It seems likely that a subset of the cells in the experiment is able to integrate incoming DNA, but does so at high efficiency. In effect, the selected marker allows this subset to be identified, and is highly likely to have integrated the other DNA species at the same time.

9.5 Algae

The unicellular green alga *Chlamydomonas reinhardtii* has been studied for many years, both as an important organism in its own right and as a model system for plants. It is particularly attractive in allowing transformation of the chloroplast and mitochondrial genomes in addition to the nuclear genome reference to 9.7 (Organelle transformation). Methods are also being developed for the manipulation of other algal species.

9.5.1 *Chlamydomonas reinhardtii*

The yeast *ARG4* gene for argininosuccinate lyase (an enzyme of arginine biosynthesis, converting argininosuccinate to arginine and fumarate) was used as a selectable marker in early work with *Chlamydomonas*. However, the *Chlamydomonas* genome is very GC rich. This means that many heterologous genes are not well expressed, unless they come from genomes that are also GC rich. The use of the yeast *ARG4* gene, therefore, was superseded by its *Chlamydomonas* homologue, *ARG7*. The most frequently used markers are listed in Table 9.7. Several require a suitable mutant background for selection. However, a number confer resistance to antibiotics or other toxic compounds and can be selected in a wild-type background. Other markers that are less widely used include genes for components of the flagellum and the *OEE1* gene, which encodes a component of Photosystem II and allows mutant cells lacking that polypeptide to grow photoautotrophically.

Although *ARS* sequences have been obtained from *Chlamydomonas*, most experiments simply use integration of DNA into the chromosome, which gives efficient transformation. Integration is usually by non-homologous recombination. This means that gene disruption (see Chapter 7) is difficult to achieve. However, RNAi can be used to reduce the level of expression of a target gene.

Transformation of *Chlamydomonas* is most efficient if protoplasts are used. These can be generated by treatment of intact cells

Table 9.7 Selectable markers commonly used in *Chlamydomonas reinhardtii*

Marker	Product	Properties
ARG7	Argininosuccinate lyase	Arginine biosynthesis (replaced *Saccharomyces cerevisiae ARG4*)
NIA1	Nitrate reductase	Growth on nitrate as sole nitrogen source (sensitivity to chlorate)
ble	Bleomycin-binding protein	Bleomycin (phleomycin, zeomycin) resistance
ALS	Acetolactate synthase	Resistance to sulphonylurea
aadA	Aminoglycoside adenyltransferase	Resistance to spectinomycin
aph7	Aminoglycoside phosphotransferase	Resistance to paromomycin, hygromycin B (also *aphVIII*)
MAA7	Tryptophan synthase	Counterselectable with 5-fluoroindole

with autolysin, a cell-wall digesting activity that is produced by gamete cells immediately prior to mating. Pre-incubation in medium lacking nitrogen can also be used to induce autolysin production. There are also mutant strains, most notably the one designated *cw15*, that lack a cell wall and hence form a 'natural' protoplast. DNA can be introduced into protoplasts by simple treatment with chemicals such as poly-L-ornithine and polyethylene glycol or zinc sulphate, but this often gives low and variable transformation rates. More efficient methods include agitation of cells in the presence of DNA and small glass beads or silicon carbide whiskers, biolistic transformation and electroporation. Transformation with *Agrobacterium* (see below) has also been reported. In general, electroporation gives the highest transformation efficiency, but agitation with glass beads is often used in preference because of its simplicity. Co-transformation is often seen.

9.5.2 Other algae

There are reasonably well-developed systems for the genetic manipulation of the green algae *Volvox carteri* and *Chlorella* sp., often using nitrate reductase or antibiotic resistance as selectable markers. Biolistic transformation has been used for both genera, and chemically induced DNA uptake by protoplasts has been used with *Chlorella*. There is considerable interest in developing effective methods for green algae of the genera *Haematococcus* and *Dunaliella*, which are of economic importance for the production of carotenoids. There are also several reports of transformation of red algae, such as *Cyanidioschyzon merolae*, and brown algae, such as the diatom *Phaeodactylum tricornutum* and the seaweed *Laminaria japonica*.

9.6 Vascular plants

A wide range of methods has been used for the transformation of plant cells. The most widely used model plant is *Arabidopsis thaliana*, though many other plants, such as tobacco, are also important as model systems. Transformation of economically important crop species, especially cereals, is also of particular interest. We will look at the different methods available, the most widely used of which is *Agrobacterium*-mediated transfer.

9.6.1 *Agrobacterium*-mediated transfer

This exploits the natural DNA-transferring properties of a soil-inhabiting bacterium, *Agrobacterium tumefaciens*. A related species, *Agrobacterium rhizogenes*, can also be used. The two species cause the diseases crown gall and hairy root respectively on plants. The species belong to the Rhizobiaceae, which include other important bacterial genera, such as *Rhizobium*. Both *Agrobacterium* species lead to disturbance of normal plant growth, causing a gall in one case

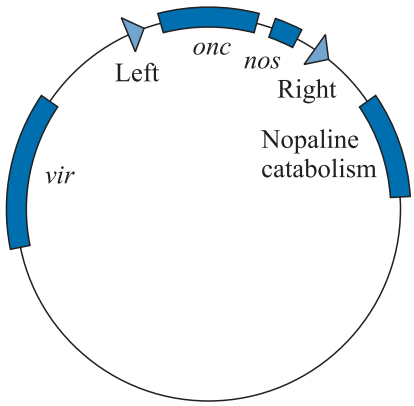


Fig 9.9 Simplified diagram of a Ti plasmid. Arrowheads indicate the left and right borders of the region transferred (the T-DNA). *Onc* genes are responsible for growth transformation, *nos* for nopaline synthase, and *vir* for transfer of the T-DNA into a target cell.

and abnormal roots in the other. Their ability to do so depends on the presence of large plasmids, in excess of 100 kbp, which are referred to as the 'tumour-inducing' (Ti) and 'root-inducing' (Ri) plasmids respectively. We will concentrate on the Ti plasmid; the Ri plasmid is quite similar.

A simplified diagram of a Ti plasmid is shown in Figure 9.9. Most important is a region referred to as the T-DNA ('transferred DNA'), which is transferred from an infecting *Agrobacterium* cell into the nucleus of the plant cell, where it is integrated into the plant genome. Transfer of the T-DNA depends on a set of genes called *vir* if they are on the Ti plasmid, or *chv* if they are on the chromosome. They are induced in response to various compounds in exudates from wounded plants; these compounds include acetosyringone, alpha-hydroxyacetosyringone and precursors of lignin, such as coniferyl alcohol. The T-DNA itself is flanked by repeated sequences of around 25 base pairs, called border repeats (or left and right borders). The T-DNA contains a group of genes referred to as the *onc* genes, as they are responsible for the T-DNA's oncogenicity (or tumour-inducing potential). They include genes for the synthesis of an auxin and a cytokinin, which are plant growth regulators that cause the disturbance of cell growth. In addition, the T-DNA contains genes for enzymes that produce unusual amino acid conjugates, such as nopaline and octopine (Figure 9.10), collectively called opines. Therefore, plant cells that have had T-DNA incorporated will grow in an apparently uncontrolled way and will synthesize opines. The opines are excreted by the plant cells (which are unable to use them themselves) and are utilized instead by *Agrobacterium* cells near the site of infection, whose Ti plasmids also contain the genes for opine breakdown. *Agrobacterium*, therefore, has a very sophisticated mechanism for subverting the normal cell function.

Vectors

The basic principle of using *Agrobacterium* as a tool for plant genetic manipulation is to insert foreign DNA into the T-DNA of a bacterial cell and rely on the bacterium to transfer the DNA into the plant.

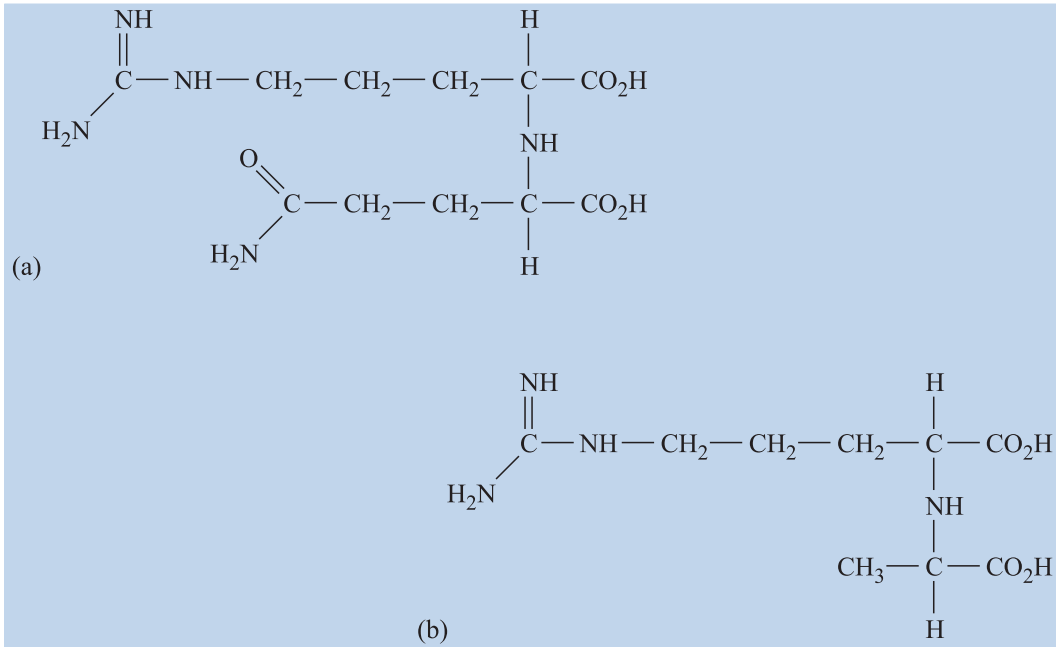


Fig 9.10 Examples of opines: (a) nopaline and (b) octopine.

As long as the necessary proteins are provided by the bacterium, any sequences flanked by the T-DNA border repeats can be transferred into the recipient plant cell genome.

There are two difficulties with this. The first is that the Ti plasmid is too large to manipulate in vitro. The second is that T-DNA contains genes that will disturb the growth of plants into which it has been transferred. These problems were initially overcome by the use of cointegrative vector systems. These used a small intermediate vector that could be manipulated conveniently and carried a selectable marker. The gene of interest was inserted into this and the recombinant intermediate vector introduced into *Agrobacterium* cells containing a Ti plasmid. The intermediate vector was not able to replicate inside *Agrobacterium*, but could be propagated by recombinational insertion into the T-DNA region of the resident Ti plasmid. (This event could be selected for using the marker on the intermediate vector.) *Agrobacterium* could then transfer the T-DNA into the recipient plant. For the recipient plants not to have abnormal growth, it was also necessary for the resident Ti plasmid to have been disabled, by inactivation of one or more of the genes for biosynthesis of growth regulators.

This approach was rather complex, because it required recombination between the integrative vector and the resident Ti plasmid in vivo in the *Agrobacterium* cell. It was superseded by the use of binary vector systems (Figure 9.11). These also use a small plasmid, the binary vector, that can be readily manipulated in vitro; but, unlike the cointegrative vector systems, this plasmid contains a pair of border repeats. The DNA of interest is inserted into the region of this plasmid between the border repeats. The recombinant binary

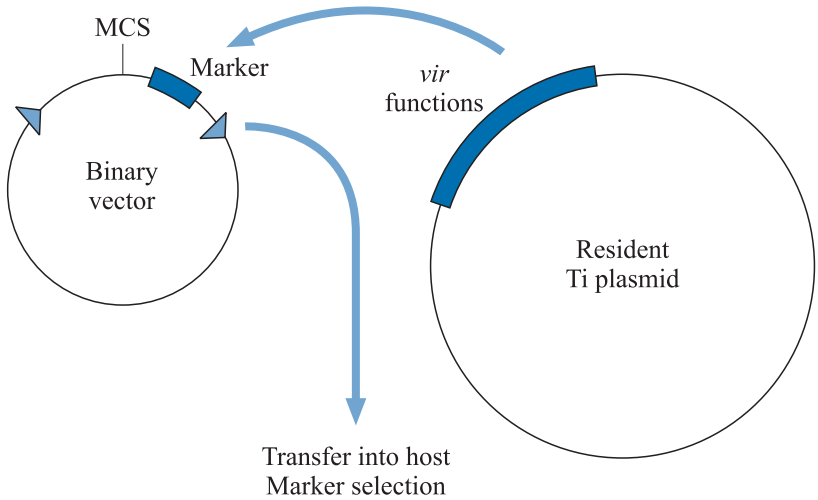


Fig 9.11 Binary vector systems. A disabled resident Ti plasmid directs transfer of T-DNA from a smaller vector (bounded by arrowheads in the figure) into the host genome. Selection is then imposed for the marker gene carried in the T-DNA.

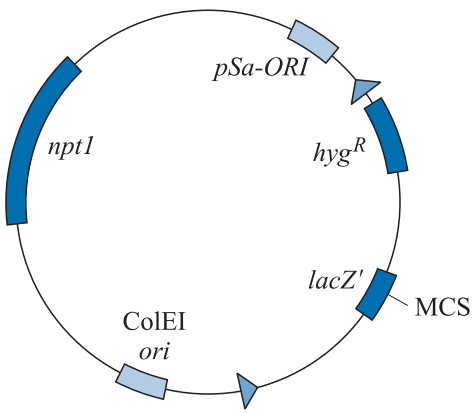


Fig 9.12 The binary vector pGreenII.000 (3.3 kb). The border repeats (arrowheads) flank T-DNA containing a *lacZ'* sequence with a multiple cloning site (MCS), and a selectable marker (*hygR*; hygromycin resistance). The pSa-ORI origin of replication and *nptI* marker allow propagation in *Agrobacterium*.

vector is then introduced into *Agrobacterium* containing a resident Ti plasmid (which has been disabled so that it cannot itself be transferred into a plant host). The *Agrobacterium* cells are then used to transfer DNA into the plant host. The resident Ti plasmid provides the functions needed (from its *vir* genes) to direct transfer of the region within the binary vector that is flanked by border repeats into the recipient plant genome.

The features of binary vectors are illustrated by the example shown in Figure 9.12. This particular one belongs to the pGreen series. It is a shuttle vector, so it can be propagated in *E. coli* or in *Agrobacterium*. In this case, there are two origins of replication. The ColEI origin allows replication in *E. coli* and the pSa-ORI allows replication in a suitable *Agrobacterium* strain. Some binary vectors use a single origin that can function in both hosts. The *nptI*

gene encodes a neomycin phosphotransferase that confers kanamycin resistance and allows the presence of the plasmid to be selected in a bacterial host. The left and right borders mark the region that will eventually be transferred to the plant host. Within this region are a multiple cloning site within a *lacZ′* minigene and a selectable marker (in this case for hygromycin resistance) allowing selection in the plant hosts. Selectable markers are discussed in more detail below. Other binary vectors include features such as additional plant promoters to drive expression and *cos* sites to allow packaging into lambda phage heads.

Once the correct sequences have been incorporated into the multiple cloning site of the binary vector, they must be transferred into a recipient *Agrobacterium* strain carrying an appropriate Ti plasmid. This is usually done by electroporation with naked DNA. However, transfer by conjugation can also be used. Conjugation often requires a triparental mating involving the *Agrobacterium* strain, an *E. coli* strain containing the vector to be transferred and an *E. coli* strain with a plasmid capable of mobilizing the binary vector into *Agrobacterium*.

Transfer into plant cells

The next stage is to arrange for the modified *Agrobacterium* to infect plant cells and transfer its T-DNA into them. Various methods exist, and the optimal ones vary from species to species. One approach (co-cultivation) is to incubate explanted material, such as leaf discs, shoots or roots, with *Agrobacterium* cultures containing the appropriate vector for a while and then transfer them to a suitable medium to select for transformation. This medium contains a selection system (according to which selectable marker was present in the vector), an additional antibiotic to kill residual *Agrobacterium* cells and plant growth regulators to assist in regeneration. Intact plants are then regenerated by tissue culture techniques, and these should contain the DNA transferred from the vector. An approach widely used with the model plant Arabidopsis is floral dipping. In this, plants are dipped briefly into a culture of *Agrobacterium* and then grown until seeds are produced. The seeds are subsequently germinated on an appropriate selective medium, and the seedlings that are able to grow should contain the DNA transferred from the binary vector. Spraying cultures onto floral tissues may also be adequate. In some protocols, plants are subjected to vacuum treatment while immersed in the bacterial culture. This is called vacuum infiltration. Other techniques for transfer of DNA from *Agrobacterium* include injection of bacteria into plant tissue and incubation of protoplasts with bacterial cultures followed by regeneration of plants from the protoplasts. In many cases the resulting plants can have several copies of the incoming DNA integrated at more or less random sites, sometimes with rearrangements, so detailed analysis of the transformed plants will be necessary.

9.6.2 Other methods of DNA transfer

Not all plant species can be successfully transformed using *Agrobacterium*. Another approach is to use biolistic transformation. This was first achieved using intact epidermal cells from onion as the target and microprojectiles carrying RNA from tobacco mosaic virus or a DNA construct encoding chloramphenicol acetyltransferase. Expression of both these nucleic acids was detected. In most applications, the gun is used to fire the microprojectiles into suitable target tissue such as leaf, after which transformed plants are regenerated on an appropriate selective growth medium.

Direct transformation of protoplasts without using *Agrobacterium* is possible. Uptake of naked DNA can be brought about by treatment with polyethylene glycol or by electroporation. DNA is integrated into the recipient genome, and in some cases at least this appears to be brought about by homologous recombination. Selection for transformants can be done using the same selectable markers as are used in *Agrobacterium* transformation. However, the most difficult aspect of this approach is the regeneration of intact plants from protoplasts. Although regeneration is possible for many plants (unlike most other eukaryotes), for many agriculturally important plants it is very difficult.

Other techniques are available, although they are less widely used than biolistic transformation or protoplast transformation. One technique uses viruses such as cauliflower mosaic virus (CaMV). Cloned DNA from CaMV is infectious when rubbed onto leaves of susceptible plants. Therefore, DNA can be inserted into the CaMV genome and then transferred into plants by rubbing onto the leaves. Although there are significant restrictions as to where DNA can be inserted into the CaMV genome to retain viability, modifications to at least one of the genes, ORF II, are possible. In one example, a chimeric molecule containing a sequence encoding a methotrexate-resistant version of the enzyme dihydrofolate reductase was introduced into turnip plants. The resulting plants produced the methotrexate-resistant enzyme and showed increased tolerance of the inhibitor when it was sprayed onto their leaves. Although this demonstrated the feasibility of the approach under certain circumstances, it has not found general application; this is partly because of the narrow constraints on the sequences that the virus can accept without loss of viability and partly because the viral DNA is presumably not integrated into the recipient, and transformation is not stable. Other methods that have been tried include direct injection of DNA, fusion of protoplasts with DNA-containing liposomes, biolistic transformation of pollen (followed by pollination with the transformed pollen) and injection of DNA into pollen tubes. These approaches are much less widely used, and in some cases there may be doubt as to whether transformation has actually taken place.

9.6.3 Transient expression

For some experiments it may not be necessary to regenerate intact plants. For example, many studies on the regulatory elements necessary for the expression of particular genes have been carried out by introducing the elements (or modifications of them) into protoplasts without stable transformation and then monitoring expression. Because the transformation is not stable, these are called transient expression systems (although, if things are working well, expression can be followed for days). When interpreting the results of these experiments, it should always be borne in mind that expression in protoplasts can be rather different from expression in the intact tissue, but for many situations the results appear quite reliable. The particle gun can be used for transient expression in target tissue that is more intact than protoplasts.

9.6.4 Promoters and selectable markers

One of the most widely used promoters is from the CaMV. It is called the 35S promoter, as it is responsible for the generation of a 35S transcript in infected cells. It is constitutive and very active, as is the *nos* promoter, which is also widely used. This comes from the nopaline synthase gene from the Ti plasmid. Although the *nos* gene is prokaryotic, it is functional in plants, where it directs the synthesis of the opine, which is exploited by other *Agrobacterium* cells. Slight adjustments have been made to the sequence to improve expression. Both the 35S promoter and the *nos* promoter have relatively little tissue specificity in their expression. Tissue specificity can be achieved by incorporation of control sequences from other promoters, such as the *rbcS* promoter (for the gene encoding the small subunit of ribulose bis-phosphate carboxylase), which is particularly active in illuminated leaves and stems. The *rbcS* promoter is also often used without association with the 35S or *nos* promoters. A very wide range of promoters is now available, to direct expression in many different tissue types. A tetracycline-regulated system (see Section 9.10.1) can also be used, directing expression in response to added tetracycline analogues.

The majority of selectable markers used confer resistance to antibiotics or herbicides. The most widely used selectable marker is resistance to kanamycin and related aminoglycoside antibiotics (such as neomycin, gentamycin, G418 and hygromycin) conferred by a bacterial aminoglycoside phosphotransferase gene. Kanamycin is particularly widely used with dicotyledonous plants; but many cereals are tolerant to this, and in such cases G418 and hygromycin are often used instead. The herbicide phosphinothricin (also known in modified forms as Basta, Liberty or bialaphos) inhibits glutamine synthetase. Two genes are available as selectable markers conferring resistance to phosphinothricin. These are the *bar* and *pat* genes, which were isolated from resistant microorganisms. They both encode a phosphinothricin acetyltransferase, which inactivates the

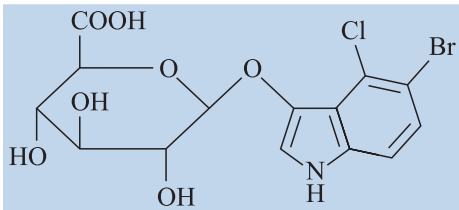


Fig 9.13 Structure of X-gluc (5-bromo-4-chloro-3-indolyl-β-D-glucuronide).

herbicide, and the genes can, therefore, be used as selectable markers when placed under the control of plant promoters. The use of markers conferring resistance to herbicides and antibiotics in transgenic crop plants has caused public concern, on the grounds that these genes may be transferred laterally. Two main approaches have been taken to deal with this. One has been to develop systems for the elimination of marker genes, in many cases by the Cre-*lox* system (see Chapter 7) once stable transformation has been achieved. The other approach has been to use markers that do not require antibiotics for selection. One strategy exploits the fact that many plants are unable to use the sugar mannose. The phosphomannose isomerase gene allows plants to grow on mannose as a carbon source as it converts mannose-6-phosphate to fructose-6-phosphate, which can be metabolized. A xylose isomerase gene has been used in a similar way, allowing growth on xylose as carbon source.

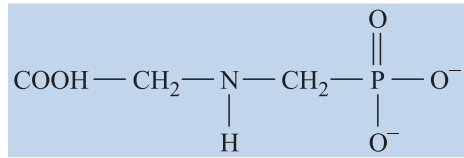
9.6.5 Reporter genes

Genes for the fluorescent proteins, such as GFP and the other fluorescence colours, are widely used as reporter genes in plants. The gene for beta-glucuronidase was used widely before the development of GFP. Beta-glucuronidase hydrolyses an artificial substrate X-Gluc (5-bromo-4-chloro-3-indolyl-β-D-glucuronide; Figure 9.13) to generate a blue pigment (analogous to the hydrolysis of X-gal by beta-galactosidase described in Chapter 3). There is very little, if any, endogenous beta-glucuronidase activity in most plant tissues, but tissue sections can be readily stained for the enzyme with X-Gluc and plant extracts assayed for it. This makes the enzyme particularly useful as a reporter. Other substrates, such as 4-methyl-umbelliferyl-β-D-glucuronide, can also be used. In this case the product is measured fluorometrically. Other reporter genes include those for luciferase, which is assayed by light production, and chloramphenicol acetyltransferase, assayed by its ability to transfer radioactive acetyl groups onto chloramphenicol.

9.6.6 Applications

Methods for the genetic manipulation of plants have, of course, been very useful for understanding plant biology. However, they have had

Fig 9.14 Structure of glyphosate.



a number of practical applications, in the development of crop plants with novel characteristics. We will look at a few particularly interesting examples.

1. **Herbicide resistance/tolerance.** This is usually achieved by causing plants to overexpress the target protein for a herbicide in order to titrate the herbicide out, to produce a protein of modified sequence that is now resistant or to produce an enzyme to detoxify the herbicide. Resistance or tolerance to the phosphinothricin herbicides can be conferred by expression of an acetyl-transferase gene. Another herbicide of great interest is glyphosate (also known as Roundup; Figure 9.14) which inhibits the enzyme 5′-enolpyruvylshikimate-3-phosphate synthase (EPSP synthase), an important enzyme of the shikimic acid pathway for the synthesis of aromatic amino acids. Glyphosate resistance can be engineered by the expression of a mutant protein that is tolerant to glyphosate and a bacterial glyphosate oxidoreductase that degrades it. A number of important crops have been engineered to glyphosate resistance, including maize, oilseed rape (canola), soya and beet.
2. **Virus resistance.** Causing plants to overexpress viral coat proteins often confers resistance to the corresponding virus, perhaps through interfering with the infection process. Tomatoes, peppers, cucumbers and papaya are among the plants that have been engineered to virus resistance in this way. Virus resistance can also be engineered using RNAi methods (see Chapter 7).
3. **Insect resistance.** Many strains of the bacterium *Bacillus thuringiensis* produce proteinaceous insecticidal toxins during sporulation, and these can be used directly as insecticides. Genes for the toxins can be incorporated into plants under appropriate promoters to cause the plants to become insecticidal themselves. Cotton was the first major crop to be engineered to insect resistance in this way.
4. **Control of ripening.** Ripening in tomatoes depends on the production and response to ethylene. This is produced from 1-aminocyclopropane-1-carboxylate (ACC) by the ethylene-forming enzyme. ACC is produced by ACC synthase (Figure 9.15). Inactivation of expression of either of these enzymes by use of antisense RNA generates plants yielding fruit that do not ripen of their own accord (as they are unable to produce ethylene),

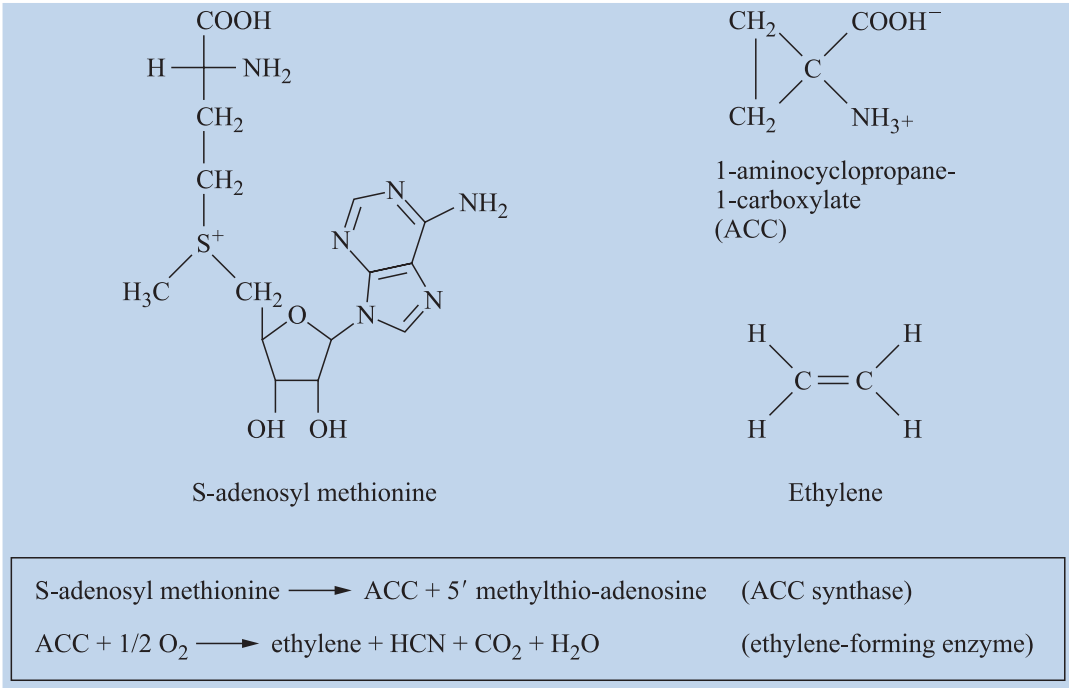


Fig 9.15 Synthesis of ethylene from *S*-adenosyl methionine via 1-aminocyclopropane-1-carboxylate (ACC).

but they will do so if given exogenous ethylene. The 'Flavr Saver™' tomato produced by Calgene contained an antisense gene for polygalacturonase, an enzyme that causes wall softening during the ripening process. The tomato was also found to have improved processing qualities.

5. **Male sterility.** Male-sterile plants (that do not produce viable pollen) are an important tool in plant breeding, to avoid self-pollination. Incorporation of a gene for a bacterial ribonuclease (barnase) attached to a promoter specific for the tapetum (tissue surrounding the pollen sac, which is important in pollen development) leads to failure of pollen formation and male sterility. This forms the basis of terminator technology used by plant breeding companies to stop the propagation by farmers (or in the event of 'escape' into the wild) of transgenic crops.
6. **Other physiological modifications.** A huge range of other physiological modifications has been achieved. These include developing resistance to chilling, drought, salinity and fungal and other pathogens. The nutritional quality of grain can also be enhanced, with important implications for developing countries. Inactivation of genes for enzymes in caffeine biosynthesis has been suggested as a way of generating 'naturally' decaffeinated coffee beans.
7. **'Pharming'.** Plants are attractive as hosts for the expression of pharmaceutical products. They have been used for the expression of vaccines, recombinant antibodies and other pharmaceuticals, such as hormones.

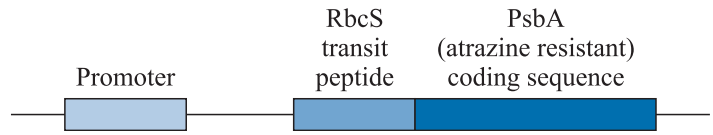


Fig 9.16 Construct used for allotopic expression of *psbA*. Details are given in the text.

9.7 Organelle transformation

The examples we have covered so far have dealt with the modification of the nuclear genome. However, eukaryotes contain one or two additional genomes, located in the chloroplast and the mitochondrion. They encode a small but significant part of the polypeptides of those organelles. The chloroplast genome encodes mainly polypeptides involved in the light reactions of photosynthesis, protein synthesis and a limited number of other functions. The mitochondrial genome encodes mainly polypeptides involved in oxidative phosphorylation. In addition, both organelle genomes encode rRNAs and tRNAs. Mutations in mtDNA frequently result in the cell's inability to carry out oxidative phosphorylation, and we will see later that this has been exploited in mitochondrial transformation work.

Early attempts to manipulate organelle genomes included the use of *Agrobacterium* and direct DNA uptake, but they met with limited success. It was difficult to demonstrate DNA uptake conclusively and reproducibly. For some purposes, a different approach could be taken by fusing to a gene a sequence that encoded an organelle-targeting peptide and then inserting the modified gene into the nucleus, rather than into the organelle. An example of this was the engineering of tolerance in tobacco to the herbicide atrazine. This herbicide acts on one of the polypeptides of the reaction centre of Photosystem II, which is the product of the chloroplast *psbA* gene. A construct was made containing a mutant *psbA* gene (whose product was resistant to atrazine) from the plant *Amaranthus* fused to a sequence encoding the transit peptide of the small subunit of ribulose bis-phosphate carboxylase. (This polypeptide is encoded in the nucleus, synthesized in the cytoplasm and subsequently imported into the chloroplast. The transit peptide is a region at the amino-terminus of the protein, which directs import into the chloroplast and is subsequently removed proteolytically.) This hybrid gene, with a plant nuclear promoter attached (Figure 9.16), was introduced into the tobacco nucleus by *Agrobacterium*-mediated transformation. The PsbA protein was produced and imported into the chloroplast under the direction of the transit peptide, and the plants showed a significant level of atrazine resistance. Similar approaches can be used to target other polypeptides to the chloroplast or the mitochondrion as appropriate, and if the presence of the normal gene in the organelle is not a problem, then inserting modified organelle genes into the nucleus may well be suitable. This is sometimes called allotopic gene expression. For some

species, direct modification of the chloroplast and mitochondrial genomes is possible.

9.7.1 Chloroplasts of algae

The first chloroplast manipulation experiments used *Chlamydomonas* strains mutant in the chloroplast gene *atpB*, encoding the beta subunit of the ATPsynthase, an important multisubunit complex required for the generation of ATP using light energy. The mutant had a partial deletion of this gene, rendering it unable to grow photosynthetically and dependent on a suitable fixed carbon source, such as acetate, in its growth medium. Cells were bombarded with a plasmid containing the cloned wild-type gene, incubated for a while, then transferred to medium lacking a fixed carbon source and illuminated, thereby selecting for the restoration of photosynthetic competence. Photosynthetic colonies were obtained, containing an apparently normal beta subunit in the ATP synthase, and with the mutant region of chloroplast DNA replaced with the wild-type region, as judged by Southern analysis. This replacement was presumably the result of homologous recombination, and all copies of the chloroplast genome contained the wild-type sequence, since restriction fragments corresponding to the mutant could not be detected. The lack of any copies of the mutant allele is remarkable because the chloroplast DNA is present at a high copy number, and it indicates that the wild-type phenotype will be stable in the absence of continued selection.

Other selectable markers (with suitable chloroplast expression sequences attached if necessary) that have been successfully used in chloroplast transformation in *Chlamydomonas* include:

(a) a bacterial *aadA* gene, encoding an aminoglycoside adenyltransferase and conferring resistance to spectinomycin or streptomycin;
(b) *aphA-6*, an aminoglycoside phosphotransferase conferring kanamycin resistance;
(c) restoration of photosynthetic competence in strains with a mutation in a gene for one of the proteins involved in the light reactions of photosynthesis;
(d) resistance to streptomycin, spectinomycin or erythromycin conferred by mutations in the 16S or 23S rRNA genes.

The selectable markers can also be used as the basis for gene disruption. Reporter genes, such as those for green fluorescent protein and firefly luciferase, are available. An advantage with the use of *Chlamydomonas* for transformation (and especially for gene disruption) is that, unlike higher plants, the organism has a single chloroplast; therefore, there is no risk of getting cells that contain a mixture of chloroplasts, some of which have been genetically altered and some of which have not. Such mixed cells are said to be heteroplasmic, rather than homoplasmic. There is still the possibility

of getting mixed populations of genomes within the single chloroplast, and this may occur if disruption of all copies of the gene would result in cell death. Biolistic transformation is the most widely used way of introducing DNA into the *Chlamydomonas* chloroplast, but chloroplast transformation by rapid mixing of protoplasts with DNA and glass beads, or electroporation, has also been reported. Chloroplast transformation systems have been established for other species, including *Euglena* and the red alga *Porphyridium*.

9.7.2 Chloroplasts of higher plants

Transformation of chloroplasts of higher plants was first reported definitively with tobacco, a couple of years after the description of transformation of *Chlamydomonas* chloroplasts. The usual approach is to use a vector that cannot replicate in the chloroplast, but contains a selectable marker and the gene of interest flanked by stretches of chloroplast DNA that are adjacent in the chloroplast genome. The vector is introduced by biolistic transformation. Homologous recombination across the chloroplast DNA sequences allows the integration of the selectable marker and the gene of interest. A few vectors that contain chloroplast origins of replication and can replicate independently of the chromosome have also been developed, and transient expression systems have also been used. The most widely used selectable markers are antibiotic resistance genes attached to appropriate plastid expression signals. They include *aadA*, encoding aminoglycoside adenyltransferase and conferring resistance to spectinomycin and streptomycin, and a variety of aminoglycoside phosphotransferase genes, conferring resistance to neomycin or kanamycin. These markers can be used for gene disruption experiments. There is also a counter-selection system, based on the cytosine deaminase gene. The product of this converts 5-fluorocytosine to the toxic 5-fluorouracil. If the gene has been introduced into the chloroplast, its subsequent loss can be selected by growth in the presence of 5-fluorocytosine. Genes for fluorescent proteins and for beta-glucuronidase are used as reporters.

Chloroplast transformation has been reported for a number of higher plants, as well as tobacco. These include rice, potato, tomato, petunia, cotton and *Arabidopsis*. It has been suggested that incorporation of transgenes into the chloroplast may offer more biological containment than when transgenes are incorporated into the nucleus. This is because in many important crop lines, the chloroplast is not transmitted through the pollen. However, it has been shown that chloroplast DNA sequences can transfer to the nucleus at a surprisingly high frequency and, therefore, evade this containment. As with manipulation of the nuclear genome, concern has also been expressed at the widespread release of antibiotic resistance marker genes when genetically modified crops are planted. This has led to the development of a Cre-lox system for removal of marker genes after stable transformation. In this,

the construct used for transformation includes a marker gene flanked by *loxP* sites. After stable transformation has been established, the transgenic line is crossed to a line that contains a nuclear gene for a chloroplast-targeted Cre recombinase. This catalyses the loss of the marker gene from the chloroplast.

9.7.3 Yeast mitochondria

Transformation of *Saccharomyces cerevisiae* mitochondria can be achieved using biolistic transformation, and was first reported in 1988. This example used yeast cells that were mutant in the nuclear *URA3* gene and in the mitochodrial *COX1* gene, encoding subunit I of cytochrome oxidase. The mitochondrial mutation rendered the cells unable to carry out oxidative phosphorylation and, consequently, unable to grow on glycerol as a carbon source (as cells cannot grow on glycerol by fermentation alone). The cells were bombarded with particles carrying a mixture of DNAs containing wild-type *COX1* and *URA3* and selected for the acquisition of wild-type *URA3* but not, initially, wild-type *COX1*. Colonies that grew in the absence of exogenous uracil were then plated onto medium containing glycerol as the sole carbon source to select for wild-type *COX1* cells, which were indeed obtained. It was argued that the initial selection for *URA3* acquisition would identify those cells that had received DNA and, from these, the cells that had incorporated the mitochondrial DNA could be selected more easily. As with chloroplast transformation, it seemed that the wild-type sequences were integrated by homologous recombination to replace the defective sequences.

It is often more convenient to introduce DNA first into mitochondria of mutant strains of *Saccharomyces cerevisiae* that have had all their endogenous DNA deleted (such strains are designated *rho*^{0}). The DNA that is introduced into the mitochondria becomes reiterated into concatemers, and the resulting individuals are often called *synthetic rho*$^-$, because they resemble naturally occurring *rho*$^-$ strains, which have had most of their mitochondrial DNA deleted and the remainder formed into a concatemer of repeating units. These synthetic *rho*$^-$ strains can subsequently be used to transfer DNA into wild-type strains, by mating. During this process, mitochondria fuse and recombination takes place between mitochondrial genomes. Generating a synthetic *rho*$^-$ intermediate and then transferring the DNA by classical genetics into a recipient strain for analysis of transgene function has a number of advantages. It does not require the construction of a new transformation host strain for each separate experiment, and transformation of rho^0 mitochondria is more efficient than transformation of mitochondria that already contain DNA. A typical strategy is summarized in Figure 9.17. The gene of interest is inserted into a vector that contains a selectable marker, typically the *COX1* gene. The vector also contains a nuclear selectable marker, usually *URA3*, although this can be supplied

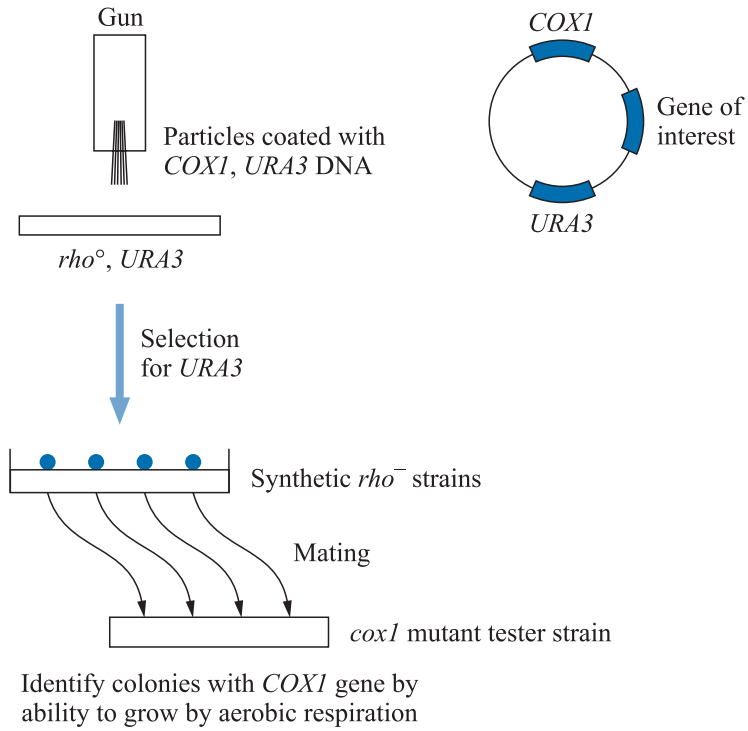


Fig 9.17 Mitochondrial transformation in *Saccharomyces cerevisiae*. Colonies that have acquired *URA3* are tested for the mitochondrial *COX1* gene by mating to a mutant strain lacking functional *cox1*. They also have the gene of interest in the mitochondrion. The gene of interest can then be transferred into other mitochondrial types by mating.

instead on a separate DNA molecule. The DNA is then introduced into a *rho*^{0} host by biolistic transformation, and nuclear transformants selected by growth in the absence of added uracil. This generates a large number of nuclear transformant colonies. Cells from each colony are then picked off and mated with a *cox1* tester strain (i.e. one that has a mutation in its *COX1* gene). If a picked cell gives rise to products in the mating that are able to grow by aerobic respiration, it must carry a wild-type *COX1* gene and, therefore, be a mitochondrial transformant. (Typically, one nuclear transformant in 1000 is also a mitochondrial transformant. Mating with the tester strain can be carried out easily. A lawn of tester strain is poured on a suitable plate and samples of the colonies to be tested are spotted directly on to it. Growth indicates that the colony tested contains the *COX1* gene.) We have, therefore, identified transformed cells that have acquired the gene of interest, and this can be transferred into other strains by conventional crossing. The three stages are therefore:

(a) selection for nuclear transformants using *URA3*;
(b) screening of those to identify mitochondrial transformants;
(c) transfer of the mutant allele into other recipient strains.

The *COX3* gene is also used as a selectable marker. It confers the ability to grow by aerobic respiration on cells with a mutation in the mitochondrial *COX3* gene. Other genes that are not of mitochondrial origin can also be used, after modification to follow the

mitochondrial genetic code. These include the *ARG8*^{m} gene for acetylornithine aminotransferase, an enzyme of arginine biosynthesis. This enzyme is usually located in the mitochondrion, but is the product of a nuclear gene. The *ARG8*^{m} gene is designed for expression in the mitochondrion and will confer the ability to grow in the absence of arginine on cells with a nuclear *arg8* mutation. The *ARG8*^{m} gene is also a convenient reporter gene. Another selectable marker is a modified form of the gene for the RNAse inhibitor Barstar. Use of this marker requires cells carrying a gene for an RNAse, Barnase, carrying a mitochondrial targeting sequence. The Barnase enters the mitochondrion and degrades mitochondrial RNA, blocking mitochondrial function. (If the Barnase is expressed at too high a level, it will also destroy cytosolic RNA.) A Barstar gene designed for expression in the mitochondrion can be used as a selectable marker in such a strain. Incorporation of the Barstar gene into the mitochondrion allows its expression. The resulting Barstar inhibits the Barnase and restores mitochondrial function.

9.7.4 *Chlamydomonas* mitochondria

The first demonstration of transformation of the mitochondrion in *Chlamydomonas reinhardtii* used a respiratory deficient strain, *dum-1*, with a deletion in the mitochondrial gene for *cytochrome b*. This made the strain unable to grow in the dark. Biolistic transformation using microprojectiles carrying partially purified mitochondrial DNA from wild-type cells led to the recovery of cells carrying the wild-type gene for *cytochrome b* and able to grow in the dark. Manipulation of the *Chlamydomonas* mitochondrial genome is not widely used at present, but has great potential for studying mitochondrial biogenesis and function in photosynthetic organisms.

9.8 *Caenorhabditis elegans*

The nematode worm *Caenorhabditis elegans* is of great importance as a model biological system. However, genetic manipulation of the organism is complex. DNA can be injected into germlime tissue or introduced by biolistic transformation, but the DNA typically forms extrachromosomal arrays of multiple copies, which do not behave stably. The incorporation of poison sequences that are toxic if overexpressed can be used to help combat this. Under some circumstances, biolistic transformation can lead to stable chromosomal integration of sequences in single copies. In a few cases, this integration is by homologous recombination, offering the possibility of gene targeting. However, the effects of inactivation of genes are most easily determined using RNAi. Double-stranded RNA can be easily introduced for this purpose simply by feeding it to animals, soaking them in a solution of double-stranded RNA, or by microinjection.

9.9 Insects

9.9.1 Cultured cells and baculoviruses

One of the important applications of insect cell transformation is the use of baculoviruses infecting cultured insect cells as expression vectors. The most commonly used baculovirus is *Autographa californica* nuclear polyhedrosis virus (AcNPV), which was isolated from the alfalfa looper caterpillar and can infect around 30 different Lepidopteran species. The virus has a double-stranded circular genome of 128 kbp. DNA replication takes place about 6 h after infection. By about 10 h after infection, new virus particles are produced by budding from the cell membrane. These are termed extracellular virus particles. Later, occluded virus particles are produced, which accumulate in the nuclei of infected cells as occlusion bodies or polyhedra. The virus particles are held together by a 29 kDa protein called the polyhedrin protein, together with another protein, p10. They serve to protect the virus particles from environmental damage after the death of the host insect and before ingestion by another insect (after which the polyhedra dissolve in the gut allowing the virus to infect surrounding cells).

The polyhedrin protein is produced at extremely high levels (up to 20% of protein synthesis), as is p10, and at this stage in the infection cycle the expression of other genes is low. If the virus is grown in cultured cells, polyhedrin and p10 are non-essential. These facts are exploited in baculovirus expression systems. The principle is that the sequence to be expressed is inserted downstream from the polyhedrin or p10 protein promoter, and the recombinant viral genome is introduced into cultured insect cells, the cell lines most widely used being derived from ovarian tissue of the fall armyworm, *Spodoptera frugiperda*, or the cabbage looper, *Trichoplusia ni*. The polyhedrin or p10 promoter then directs high levels of expression of the inserted sequence. Other approaches have also been taken, e.g. using a nuclear polyhedrosis virus from *Bombyx mori*, then infecting silkworm larvae with the recombinant viruses and recovering protein from the haemolymph. Expression of proteins using nuclear polyhedrosis viruses is advantageous not only because it gives high protein yields, but also because a range of post-translational modifications may be carried out that are carried out much less reliably (if at all) with prokaryotic or yeast systems. These modifications include glycosylation (O- and N-linked), phosphorylation, acylation, proteolytic processing and secretion. One or more of these modifications may be required for the expressed protein to have biological activity.

However, the size of the viral genome presents a problem, as it is too large to be easily manipulated in vitro. A convenient solution is to use a system akin to the co-integrative vector system used with *Agrobacterium tumefaciens*. The baculovirus system requires a small plasmid termed the transfer vector. Figure 9.18 illustrates an example of a typical transfer vector that contains

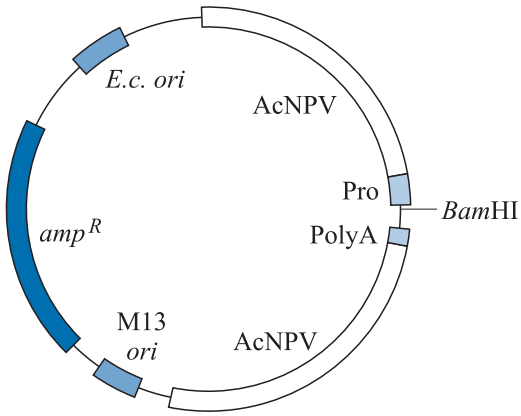


Fig 9.18 Plasmid pBacPAK1 (5.5 kb) from a baculovirus transfer vector system. Sequences can be inserted into the *Bam*HI site between the polyhedrin promoter (Pro) and the polyadenylation site (PolyA). Viral sequences on each side (AcNPV) allow integration into a complete viral genome by recombination. The plasmid can be propagated in *E. coli* using the origin (*E. c. ori*) and ampicillin resistance marker (amp^R), and there is also an origin for single-stranded DNA synthesis if required (M13 *ori*).

the following: sequences to allow propagation in *E. coli*; a viral promoter (in this case the polyhedrin gene promoter); the polyhedrin mRNA polyadenylation signal; and sequences that, in the virus, flanked both ends of the polyhedrin gene. The gene to be expressed is inserted adjacent to the viral promoter (which can be done to generate fusion proteins or intact proteins, depending on the vector used). The construct is then transfected into cultured insect cells, along with complete intact viral genomic DNA, giving rise to infection plaques in the host cells. The transfection can easily be accomplished by incorporation of the DNA into liposomes that fuse with the insect cells, or by co-precipitation of the DNA onto the cells with calcium phosphate. Once the DNA is inside the insect cells, homologous recombination can take place, in which the polyhedrin gene from the intact viral genome is replaced with the region containing the foreign DNA. This recombination occurs between the sequences flanking the polyhedrin gene in the intact viral DNA and the corresponding homologous sequences in the transfer vector. Recombination generates a modified virus containing the foreign DNA. Plaques produced by these modified viruses can be distinguished visually from plaques formed by the wild-type virus. Large quantities of cells can then be grown and infected with the modified virus. Within these infected cells, the viral promoter directs high levels of transcription of the inserted sequence. The presence of the polyadenylation signal ensures that the mRNA produced is polyadenylated. Large amounts of protein are synthesized, which can be harvested. The whole procedure is summarized in Figure 9.19.

Modifications

One potential difficulty with the system just described is the reliance on recombination to incorporate the modified sequences into the viral genome. Various modifications have been made to alleviate this.

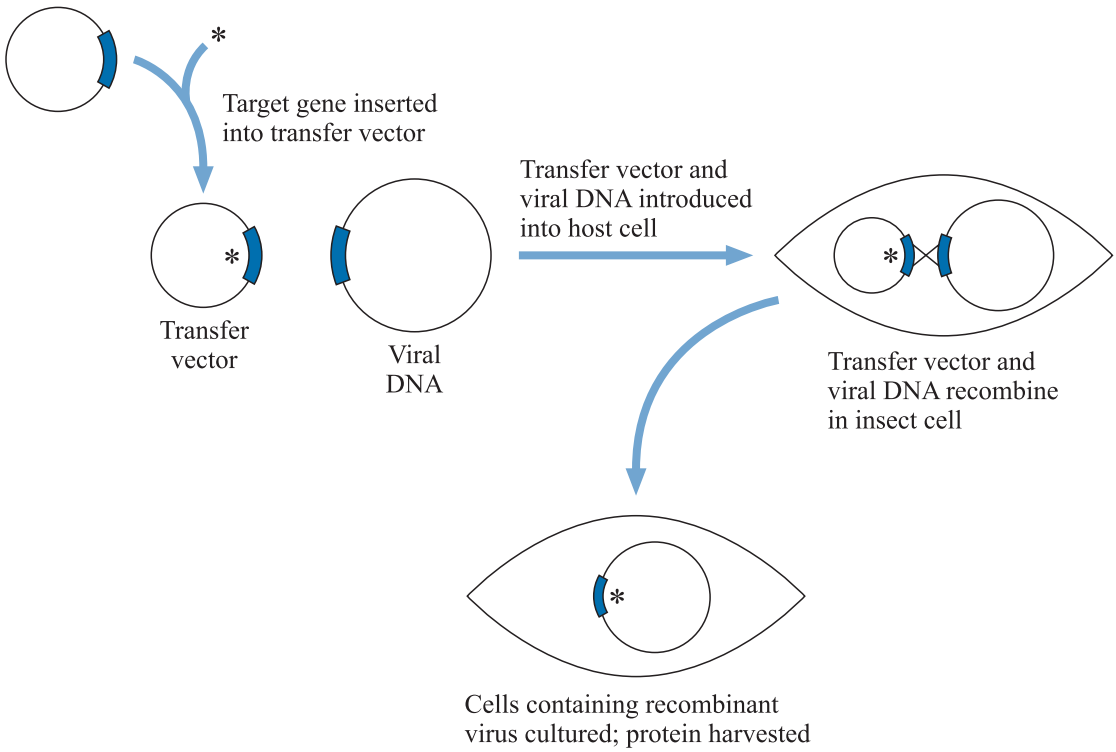


Fig 9.19 Generation of recombinant baculovirus in a host cell. The asterisk represents the gene to be expressed.

Some transfer vectors contain a *lacZ* gene adjacent to the site for insertion of the sequences for expression. The expression in plaques of *lacZ* (detected using X-gal), therefore, indicates that the plaques contain recombinant viruses. This helps in detection of recombinants, but it does not increase their frequency. Modifications to improve the recombination frequency include simply linearizing the DNA, on the grounds that this appears to make it more recombinogenic, and selection schemes whereby regions essential for growth are removed from the viral genome and can only be reconstituted by recombination with the transfer vector. Thus, the only plaques formed should contain the recombinant viruses (as only recombinant viruses have the complete genome needed to infect cells).

A more sophisticated approach allows for the generation of the recombinant virus in *E. coli*. The method uses a virus that has been modified by the addition of a replication origin and selectable marker, so it can be propagated in *E. coli*. In effect, it can function as a shuttle vector, and is sometimes known as a bacmid. Both the bacmid and the transfer (also called the donor) vector contain recognition sites for the recombination system of a transposon such as Tn7 (Figure 9.20). The transfer vector is introduced into an *E. coli* strain that carries the bacmid and a third (helper) plasmid, which encodes the transposase. The transposase catalyses the insertion of sequences from the transfer plasmid into the bacmid, which disrupts a *lacZ′* minigene in the bacmid. *E. coli* colonies containing a recombinant virus can, therefore, be identified by blue-white selection, and the viral DNA recovered

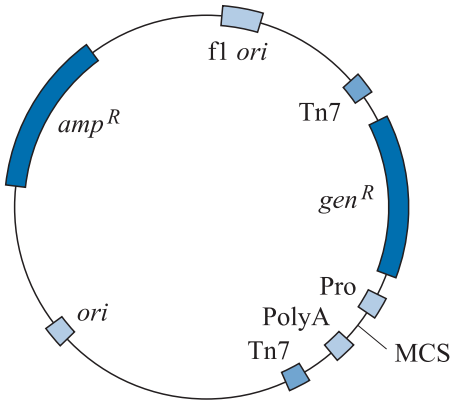


Fig 9.20 The pFastBac™ I transfer plasmid (4.8 kb). Sequences are inserted into the multiple cloning site (MCS) between a polyhedrin promoter (Pro) and an SV40 polyadenylation sequence (PolyA). The region is flanked by Tn7 transposition sequences (Tn7). The plasmid also contains genes for resistance to gentamycin (*gen*^{R}) and ampicillin (*amp*^{R}) and an origin (*ori*) for replication in *E. coli*. There is also an f1 phage origin of replication (f1 *ori*).

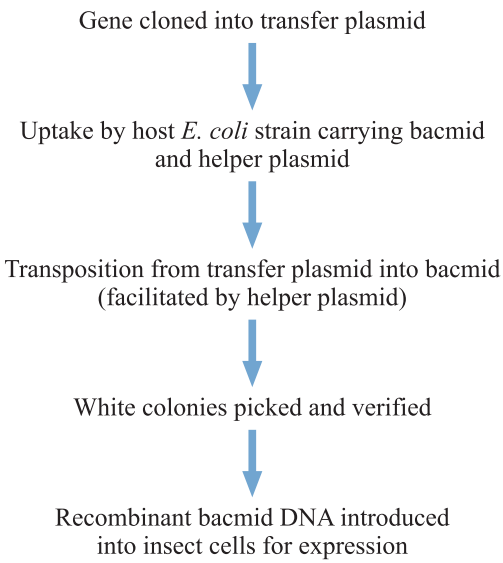


Fig 9.21 Summary of the Bac-to-Bac® cloning system for generating recombinant baculovirus.

and introduced into insect cells. Carrying out the recombination and selection in *E. coli* in this way is faster than using an insect cell host. The procedure is summarized in Figure 9.21.

Enhanced vectors

Various modifications have been made to the basic baculovirus vectors. They allow co-expression of several polypeptides, by the use of several promoters, and secretion of expressed polypeptides under the control of the honeybee melittin secretion signal.

9.9.2 *Drosophila*

Transgenic *Drosophila* can be made by microinjection of DNA into embryos. The embryos are injected at the pre-blastoderm stage, when they contain a layer of nuclei that have not been separated into individual cells. This is called a syncytium, and injecting embryos at this stage means that many nuclei will be accessible to the incoming

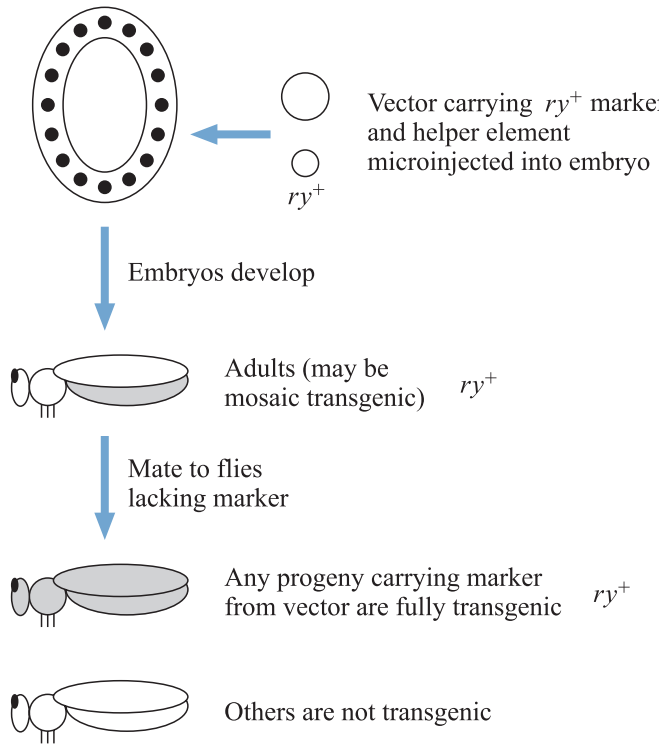


Fig 9.22 Genetic manipulation of *Drosophila*. Mosaic transgenic flies are generated by insertion of an artificial P element into germ-line cells. Progeny derived from transgenic germ cells are fully transgenic. The rosy marker is indicated by ry^+.

DNA. For reasons that will become clear, though, it is only the nuclei that will give rise to germline cells (located at one end of the embryo) that acquire DNA stably. Embryos are then grown to maturity. The mature flies will be mosaic, with some of the germline cells transgenic. Mating these flies produces progeny that are transgenic but no longer mosaic.

In contrast to what happens in many other organisms, injected DNA does not integrate efficiently into the chromosomes of *Drosophila* unaided. Therefore, it is necessary to rely on a 'natural' integration system that depends on transposable elements. For *Drosophila*, the most widely used transposable elements are called P elements, which are the basis of a phenomenon known as hybrid dysgenesis. Under the action of a transposase that is also encoded by the P element, they will insert themselves at random into the DNA of *Drosophila* germline cells. An artificial P element that does not encode transposase, but does have its site of action, can also become integrated if the transposase is provided by a helper P element also present. (This is reminiscent of the binary vector systems used with *Agrobacterium*.) So, in genetic manipulation of *Drosophila* (Figure 9.22), the sequences to be inserted are first incorporated into a modified P element, which needs to contain only a genetic marker and those *cis*-acting sequences needed for transposase action. This element is then micro-injected into pre-blastoderm embryos, together with DNA from a helper P element. This helper element directs the synthesis of the transposase, but lacks the sites for transposase action, so it

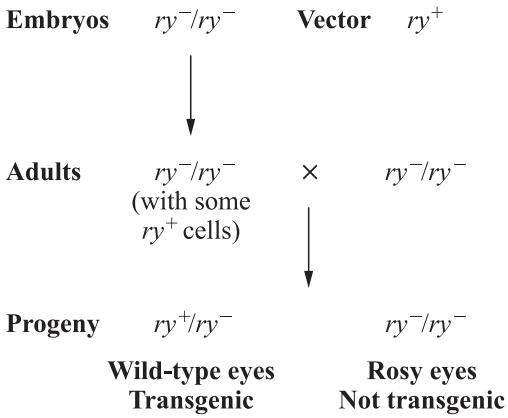


Fig 9.23 Use of the *rosy* gene in manipulation of *Drosophila*.

cannot itself transpose. Microinjection is carried out using a micromanipulator and tiny glass pipettes, although particle bombardment can also be used. Within the syncytial embryo, some of the P element DNAs will find themselves in the region that will ultimately develop into the germline. During that development, the transposase produced from the helper element causes the integration of the modified P element into the germline genome. Transgenic gametes will be produced when the fly is mature, but note that not all the germline cells will necessarily be transgenic, and the rest of the fly will not be — it is a mosaic. To generate a fully transgenic fly, we have to mate the mosaic and obtain progeny flies. Progeny flies that are produced from a non-transgenic gamete from the mosaic will not be transgenic. Progeny formed from a transgenic gamete will be fully transgenic and not mosaic. These flies are identifiable by the presence of the genetic marker on the modified P element. The site of integration of the P element seems to be more or less random, as with *Agrobacterium* T-DNA insertion. Insertion is due to a specific recombinase activity, rather than general homologous recombination.

1. **Markers.** The first marker used for transformation was the wild-type *rosy* gene, ry^+. This encodes xanthine dehydrogenase, which carries out a step in the synthesis of eye pigments. Wild-type flies have brick-red eyes, whereas *rosy* mutants have dark crimson to brown eyes. So, if the host fly embryo which is injected and the fly it is subsequently mated with are both ry^- and the vector carries the wild-type ry^+ gene, transgenic progeny flies can be identified as those with the wild-type eye colour, as shown in Figure 9.23. A disadvantage of *rosy* is that the gene is rather large. Other eye-colour markers are also commonly used, such as *white* and *vermilion*. The *rough* marker is also used, which affects eye morphology. Markers affecting other aspects of *Drosophila* biology are also available, such as the alcohol dehydrogenase adh^+ gene. Wild-type flies can be selected as they are resistant to ethanol supplied in the food, which is toxic to adh^- mutants. A bacterial

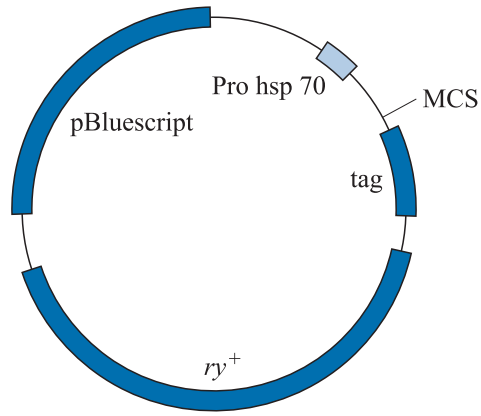


Fig 9.24 *Drosophila* pGS vector series. The vectors are constructed using pBluescript (allowing replication in *E. coli*). The example shown contains an hsp70 promoter (Pro hsp70), a multiple cloning site (MCS) allowing fusion to a tag and the rosy (ry^+) marker.

neomycin phosphotransferase gene attached to a *Drosophila* heat shock promoter is also available. It can be selected by inclusion of the antibiotics G418 or neomycin in the food.

2. **Vectors.** As well as a selectable marker (see above), a vector should contain the appropriate *cis*-acting sites for the transposition process. The promoter from the *Drosophila* heat shock gene *hsp70* is used in many vectors to direct expression of inserted sequences. There are also more complex vectors, which offer epitope tagging and other features (e.g. Figure 9.24).
3. **Helpers and hosts.** Various helper elements are available. The helper elements must encode a functional transposase, and some have been modified to try to increase its production. Helper elements generally have mutations of the *cis*-acting transposition sequences, so that although they can direct the transposition of other sequences, they cannot be incorporated into the transgenic flies themselves. This would lead to genetic instability, as the elements transposed in subsequent generations. The host strain should not contain any integrated P elements, because these might also cause instability. The host must carry suitable mutations for the subsequent detection of incorporated DNA. Tissue-specific expression of inserted sequences can be directed by generating two transgenic lines. One line contains a gene for a transcriptional activator (such as *GAL4*) under the control of a tissue-specific promoter/enhancer. The other line contains the target gene for expression under the control of the recognition sequence for the same transcriptional activator. Crossing the two lines generates progeny with the full system for expression: transcriptional activator, recognition site and target gene. This approach is particularly useful when general overexpression of the target gene is deleterious.

9.9.3 Gene inactivation in *Drosophila*

Drosophila has been an important subject for large-scale studies of the effects of systematically silencing individual genes. Because

P elements integrate at random into the genome, it is not feasible to use them directly for targeted gene disruption. However, the ability of P elements to insert into the genome has allowed the establishment of a large range of lines with genes randomly inactivated by transposon insertion. The recombinogenic activity of linear extrachromosomal DNA has also been exploited to develop recombination-mediated strategies for gene targeting. The most widely used method for gene inactivation is RNAi. Double-stranded RNA can simply be microinjected into embryos (or other developmental stages) for transient gene silencing. For long-term, or heritable, silencing, a transgene is constructed that will direct the synthesis of an inverted repeated copy of the target gene. The transgene is introduced into the target organism and the RNA transcribed from it forms a hairpin that leads to silencing.

9.9.4 Other insects

Similar transposon-based systems using transformation with microinjection have been developed for other insects. Although P element transposition is less efficient in many insects than in *Drosophila*, other transposons (especially the Hermes, mariner and piggyBac elements) have been used. Rather than relying on isolating eye-colour mutants in other species, and using the wild-type genes as selectable markers, most studies have used an artificial gene comprising a coding region for a modified GFP under the control of a synthetic promoter. This promoter comprises three copies of the binding site for a transcription factor, Pax-6, adjacent to a TATA sequence, and is referred to as the 3xP3 promoter. It functions in a very wide range of insects. GFP expression can be detected in the eyes of transgenic insects.

9.10 Mammals

The other animal group we will look at in detail is mammals, although many of the techniques outlined here can be applied to other vertebrates, such as amphibians and fish. There are three kinds of genetic manipulation that we might need to carry out. They involve cultured cells, restricted areas of intact organisms or whole individuals.

9.10.1 Cultured cells

Selectable markers

The first transformations in mammalian systems were achieved with cultured cells from humans suffering from the Lesch–Nyhan syndrome, which is caused by a deficiency of the enzyme hypoxanthine-guanine phosphoribosyl transferase (HGPRT). Cells deficient in HGPRT die in a selective medium containing hypoxanthine, aminopterin, and thymidine (called HAT medium; Figure 9.25). This is because aminopterin blocks the pathway for

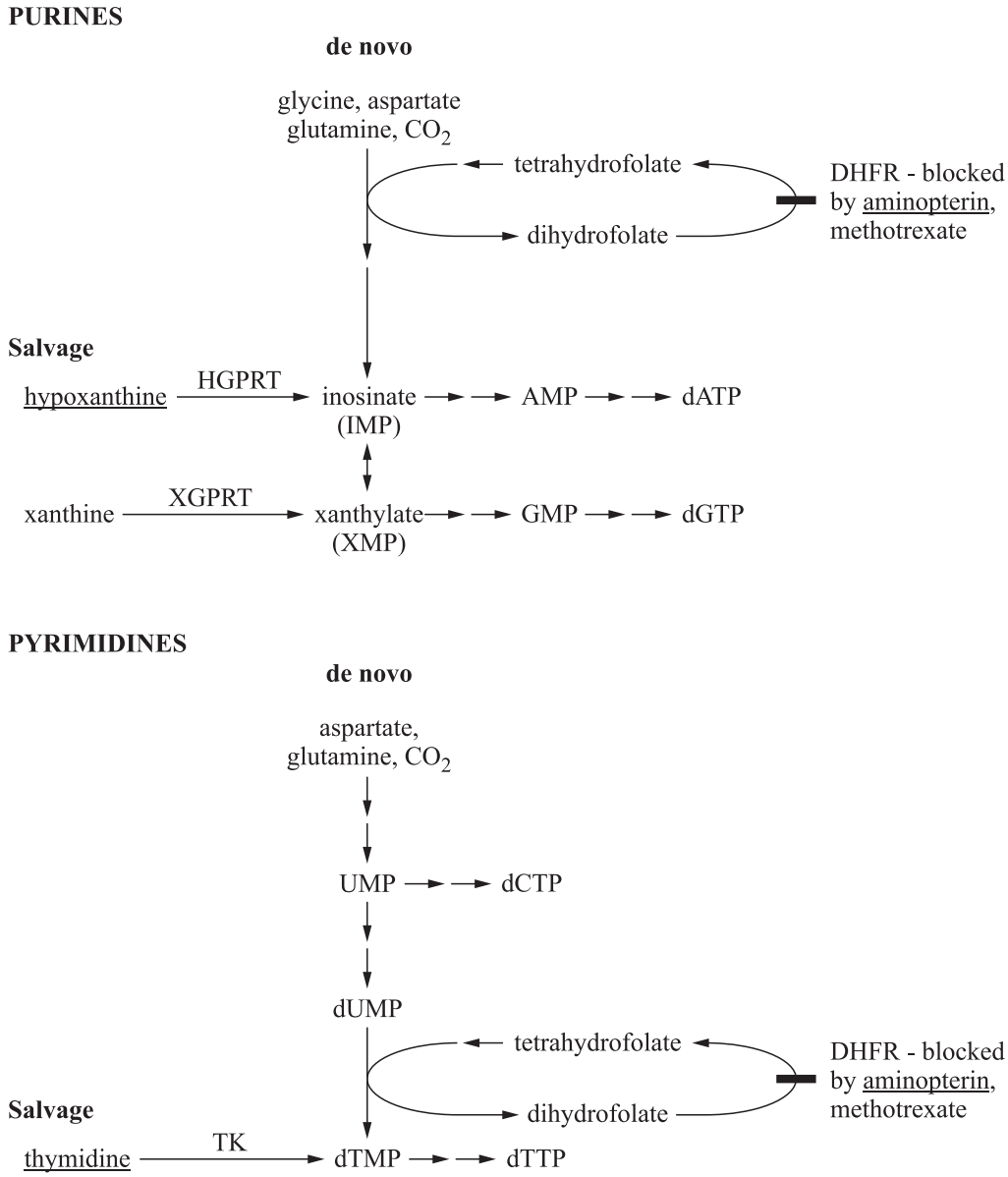


Fig 9.25 HAT selection. Compounds present in HAT medium are underlined. Aminopterin and methotrexate block the *de novo* synthesis of purine and pyrimidine nucleotides; synthesis from hypoxanthine and from thymidine require functional HGPRT and TK enzymes respectively.

endogenous synthesis of the purines needed for synthesis of nucleic acids. An alternative pathway allows synthesis of purines from hypoxanthine, but this requires the enzyme HGPRT, and so cannot operate in $HGPRT^-$ cells. So $HGPRT^-$ cells are unable to make nucleic acids in HAT medium and will die; wild-type cells are able to make nucleic acids and will live. Precipitation of wild-type DNA onto the $HGPRT^-$ cells in the presence of calcium phosphate led to DNA uptake and stable transformation. Similar results could be obtained with cells deficient in thymidine kinase (TK). They are also killed in HAT medium. Pyrimidine synthesis is also blocked by aminopterin, and utilization of the thymidine supplied in the HAT medium

(Figure 9.25) requires a functional TK. Transformation of TK^- cells with DNA containing a functional TK gene gave rise to cells that were able to survive in HAT medium.

So, suitable treatment of cultured cells can lead to uptake and expression of exogenous DNA. In addition to HGPRT and TK genes, commonly used selectable markers for mammalian cell transformation include those listed below. Note that selectable markers derived from prokaryotes have been modified to allow expression in mammalian cells.

(a) Resistance to aminoglycoside antibiotics, conferred by a bacterial aminoglycoside phosphotransferase (also known as neomycin phosphotransferase) gene. The aminoglycoside G418 is widely used for selection.
(b) Resistance to the protein synthesis inhibitor hygromycin, conferred by a bacterial *hph* gene encoding hygromycin phosphotransferase.
(c) Resistance to the protein synthesis inhibitor puromycin, conferred by a *Streptomyces* gene encoding puromycin-*N*-acetyltransferase.
(d) Resistance to the DNA damaging agent bleomycin (zeocin), by expression of a binding protein.
(e) Resistance to methotrexate, which inhibits the enzyme dihydrofolate reductase (DHFR). This enzyme is involved in metabolism of one-carbon units and is required for nucleoside biosynthesis. Resistance can be conferred (in the absence of exogenous nucleosides) by expression of high levels of a sensitive enzyme or by expression of a resistant enzyme.
(f) Resistance to toxic adenosine nucleosides such as 9-β-D-xylofuranosyl adenine. This is conferred by overexpression of adenosine deaminase (ADA) in cells that lack it, or overexpression (in the presence of higher concentrations of 9-β-D-xylofuranosyl adenine) in cells that already contain some ADA.
(g) Resistance to difluoromethylornithine, conferred by high levels of expression of ornithine decarboxylase, the enzyme it inhibits.
(h) A bacterial xanthine-guanine phosphoribosyl transferase (XGPRT) gene. Expression of XGPRT allows $HGPRT^-$ cells (or wild-type cells in the presence of aminopterin) to utilize exogenous xanthine for purine nucleotide biosynthesis.
(i) Histidinol dehydrogenase, which allows cells to synthesize histidine from exogenous histidinol, and protects against toxic effects of histidinol.
(j) Overexpressing glutamine synthetase genes, which allow cells to continue to make glutamine from exogenous glutamate and ammonia in the presence of the glutamine synthetase inhibitor methionine sulphoximine.
(k) Resistance to the peptidyl nucleoside analogue blasticidin, an inhibitor of protein synthesis, conferred by genes that inactivate the antibiotic.
(l) Cytosine deaminase.

As in some other systems, co-transformation is often observed, where cells that have been selected for acquisition of one marker are frequently found to have acquired an unlinked and unselected marker that was supplied at the same time. This means that a selected marker does not necessarily have to be on the same piece of DNA as other transforming sequences. Selection for a particular marker gene can sometimes result in amplification of the number of copies of it in the target cells, which may be useful, although the amplification is not always stable in the absence of continued selection. Counter-selection, for the absence of particular markers, is also possible. The most commonly used counterselectable markers are:

(a) TK from viruses such as herpes simplex virus. This enzyme will metabolize certain nucleoside analogues, such as gancyclovir and bromodeoxyuridine, to form products that damage DNA or block its synthesis. Although many mammalian cells contain TK, the mammalian enzyme uses the analogues much less efficiently than does the viral enzyme. Cells lacking the viral enzyme, therefore, are resistant to the analogues and are able to grow even if they contain the endogenous mammalian enzyme.

(b) Cytosine deaminase (CDA). This enzyme converts 5-fluorocytosine to the toxic 5-fluorouracil. Cells lacking the gene for CDA, therefore, are able to grow in the presence of 5-fluorocytosine. (Positive selection for the presence of CDA is also possible, by growth in the presence of *N*-(phosphonoacetyl)-L-aspartate, which blocks *de novo* pyrimidine biosynthesis by inhibition of aspartate transcarbamoylase, and cytosine, which can provide pyrimidines if CDA is present.)

(c) Diphtheria toxin, which blocks protein synthesis by ADP-ribosylation of elongation factor 2.

Hosts and transformation methods

A wide range of host cell lines is available. Commonly used human cell lines include HeLa, 293T (derived from kidney), and Jurkat (lymphoblasts from a leukaemia patient) cells. Several other tumour cell lines are also quite widely used. Cultured cell lines from other species include COS (derived from kidney cells from African green monkeys), CHO (epithelial-like cells from Chinese hamster ovary), and 3T3 (fibroblasts from mouse) cells. Embryonic stem cells are particularly important for the generation of transgenic whole organisms (see below).

The most commonly used methods for introduction of DNA are calcium phosphate co-precipitation, treatment with DEAE-dextran, liposome packaging and electroporation. The last method (and in some cases DEAE-dextran treatment) is particularly useful for cells growing in suspension. The other methods are best suited for cells growing on a surface. Calcium phosphate co-precipitation is easily achieved by mixing a solution containing the DNA with solutions containing phosphate ions and calcium ions. This results in the

formation of a very fine precipitate of calcium phosphate, particles of which are probably taken up by the cells by endocytosis, with the DNA included passively in the process. Further treatment with glycerol or dimethyl sulphoxide may improve the efficiency of uptake. DEAE-dextran is a modified polysaccharide that is positively charged. This forms a complex with the negatively charged DNA molecules and the complex is taken into the cell by endocytosis. Liposome methods generally involve mixing the DNA with one or more lipids, at least one of which is positively charged. This forms DNA-containing vesicles that fuse with the cell membrane. A wide range of transformation reagents, most of which are based on cationic lipids, are available commercially. Viral packaging, *E. coli* spheroplasts, red blood cell ghosts (cells whose contents have been removed and replaced, by swelling and shrinking in solutions of suitable osmotic strength) and microprojectile bombardment have also been used to introduce DNA into mammalian cells.

Vectors

There are very many vectors available. Some will replicate effectively in cells, allowing extrachromosomal (episomal) maintenance. Others will not and must, therefore, be used in transient expression systems unless they can integrate into the chromosome. There is a range of expression vectors, although they will of course be unnecessary if one wishes simply to test the activity of particular promoter constructs in transformed cells. Overexpression of cloned sequences, on the other hand, can be used for many purposes, including tests of biological function and library screening. Selection for complementation in cells from lines deficient in a particular function allows the identification of cells containing the equivalent cloned gene or cDNA. This has been used, for example, to recover cDNAs for genes involved in DNA repair by complementation of repair-deficient cell lines. If an overexpressed protein is expressed at the cell surface, then it may be possible to select appropriate cells using some kind of panning affinity technique using antibodies or a ligand (see Chapter 6). Clearly, the details of the selection methods used will depend on the particular function of the protein in question and the nature of the cells expressing it. There are also vectors designed for a range of specific purposes, such as two-hybrid screening, targeting of expressed proteins to specific subcellular locations, epitope tagging and reporter systems.

Many vectors are based on the virus SV40. This can replicate successfully in certain cell types, such as those derived from the African green monkey, which are said to be permissive. Additional virus particles can be produced which will go on and infect other cells, although eventually lysis will take place. In non-permissive cells, virus replication does not take place, although the viral DNA can be expressed. If selection for the incoming sequences is imposed, then the only way they can be maintained in non-permissive cells is by integration.

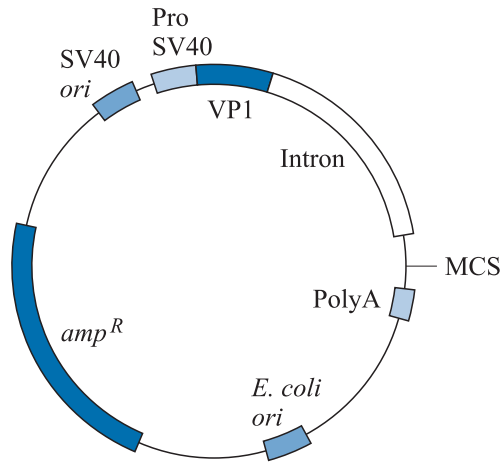


Fig 9.26 Plasmid pSVL (4.9 kb). The vector contains the SV40 late promoter (Pro SV40) and part of the region encoding a viral polypeptide (VPI), with an intron and polyadenylation site (PolyA) associated with VPI. There are SV40 and prokaryotic origins, a multiple cloning site and a prokaryotic selectable marker (*amp*^{R}).

SV40 normally produces two sets of transcripts, denoted early and late. They are initiated divergently from near the origin of replication, and production of the early transcripts is stimulated by two 72-base-pair enhancer sequences. Both sets of transcripts are spliced and polyadenylated. It seems that splicing of transcripts is necessary for efficient expression. Transcripts that have not been through a splicing process are not expressed efficiently, even if introns have been removed from the DNA. So, if vectors are based (as in many examples) upon insertion of sequences into the regions covered by the early or late transcripts, it is important not to disrupt the splicing sites, even if the inserted sequences do not themselves contain introns. An example of such a vector is pSVL, shown in Figure 9.26. This shuttle vector contains pBR322 sequences (from which a 'poison' sequence causing reduced replication efficiency has been deleted) that include the ampicillin resistance gene and the origin of replication. SV40 sequences include: the origin of replication; the late promoter; a region, transcribed from that promoter, that includes intron material; and the polyadenylation site. DNA is inserted after the intron, allowing efficient splicing and polyadenylation. Replication would be possible in permissive cells, giving a high copy number, were it not for the fact that the vector lacks the coding sequence for the T protein (a product of early viral transcription), which is also required for replication. So, in the absence of T, maintenance of the plasmid is possible only by integration. Replication in permissive cells and maintenance extrachromosomally is possible if they are co-infected with a helper virus, or if the cells produce the T protein themselves. This is the case with COS cells (described earlier). These cells will support a high copy number (and, in principle, high expression levels) of SV40-based vectors, although in some cases the copy number is not maintained indefinitely. Control of copy number is possible using a variant of COS cells that produce a temperature-sensitive T protein. Complementation to produce more virus particles (rather than just replication of viral

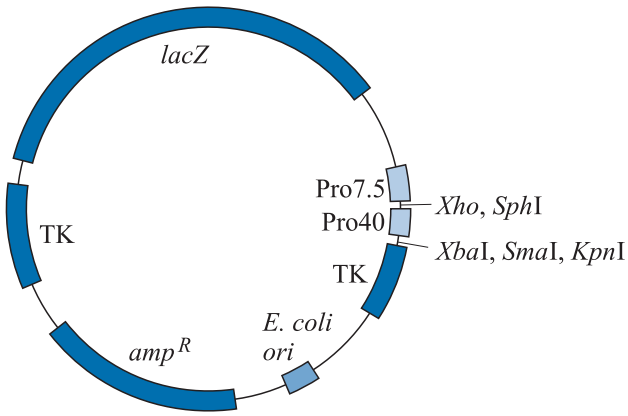


Fig 9.27 Plasmid pAbT4586 (7.8 kb), containing two vaccinia virus promoters (Pro 7.5 and Pro 40). The viral thymidine kinase (TK) sequences allow integration into a viral genome by recombination.

DNA) would require, in addition, the provision of the late viral VP1, VP2 and VP3 proteins, which can be provided by a helper virus. Packaging also places constraints on the size of insert DNA that can be accommodated.

Another widely used virus is the vaccinia virus, a member of the pox virus family. It has a large genome, nearly 200 kb, but can accept at least 25 kb of extra foreign DNA. The virus encodes its own RNA polymerase, which is functional in the cytoplasm of infected cells, where the virus also replicates. Because of the size of the viral genome, a commonly used approach is to generate recombinants in vivo in infected cells by recombination between a suitable plasmid, into which the sequences of interest have been inserted, and a parental virus strain (as with the baculoviruses). Plasmid is transfected into cells carrying the virus and recombination integrates the plasmid into the virus. Techniques are available for the selection of recombinant viruses, e.g. by conferring the ability to replicate in suitable cell lines and/or in the presence of selective agents. Recombinant vaccinia viruses can be used as live vaccines, as well as for the expression of genes in cultured cell lines.

Figure 9.27 gives an example of a plasmid vector, pAbT4586, for use in the vaccinia system. This contains two viral promoters (Pro7.5K and Pro40K) with cloning sites adjacent to them, a modified *lacZ* marker gene and a non-functional viral TK sequence. The plasmid integrates into the viral genome by recombination across the plasmid and viral TK sequences, rendering the virus TK^- because the endogenous viral TK gene is disrupted. TK^- host cells are used, which can grow in the presence of 5-bromodeoxyuridine. Host cells carrying non-recombinant virus will be TK^+, and will be killed by 5-bromodeoxyuridine. Cells carrying a recombinant virus will still be TK^- and will survive in the presence of 5-bromodeoxyuridine. Virus plaques should be blue in the presence of X-gal, because of the beta-galactosidase gene.

Retroviruses are also useful for genetic manipulation of cells in culture. They have the advantages that they will infect susceptible cells with very high efficiency and bring about stable integration into

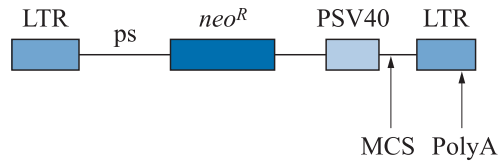


Fig 9.28 Part of the retrovirus-based vector LNSX. (For simplicity, only the retroviral part of the vector is shown.) LTR = long terminal repeat; ps = packaging signal; *neo*^{R} is a neomycin resistance gene; PSV40 is a promoter from SV40.

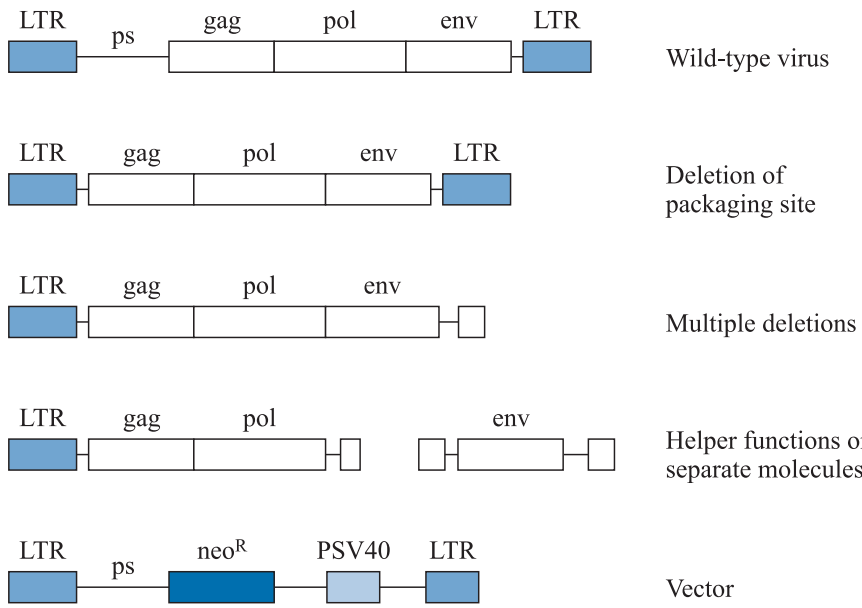


Fig 9.29 Strategies for avoiding packaging of helper retrovirus. The top and bottom diagrams show wild-type virus and a typical vector molecule (see Figure 9.28) respectively. The middle diagrams show three helper retrovirus systems that can provide all the functions for vector packaging, but cannot themselves be packaged. *Gag*, *pol* and *env* are retroviral genes; ps = packaging signal.

the target cell genome. An example, LNSX, is shown in Figure 9.28. This retrovirus contains a neomycin resistance gene as a selectable marker and a promoter (in this case the SV40 early promoter) adjacent to a cloning site. It is usually combined with a prokaryotic plasmid to create a shuttle vector. It does not encode viral proteins, so it cannot itself integrate directly into a host genome. After any construction work in a prokaryotic host has been completed, the retroviral vector is transfected as naked DNA into mammalian cells (*packaging* cells) infected with a helper retrovirus. In these cells, RNA transcripts are produced from the incoming retroviral vector DNA, and these are then packaged into retroviral particles containing the proteins (encoded by the helper retrovirus), such as reverse transcriptase, needed for subsequent infection of target cells. The resulting packaged viruses can then be used to infect other target cells efficiently, resulting in reverse transcription of the retroviral RNA and stable integration of DNA into the recipient cell genome. It is important to stop the helper

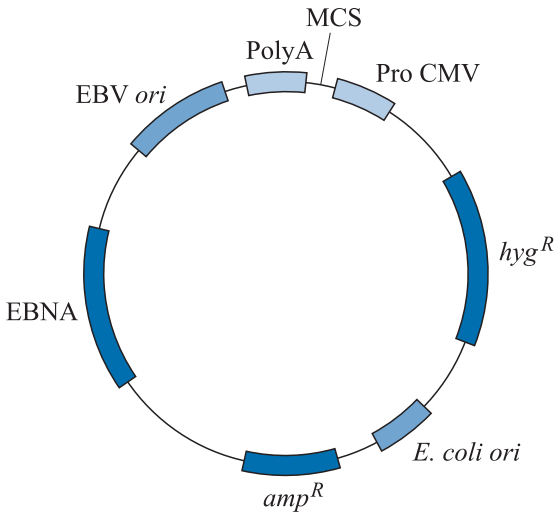


Fig 9.30 Plasmid pCEP4 (10.4 kb), based on Epstein–Barr virus. The plasmid contains an origin of replication from the virus (EBV*ori*) that, together with the nuclear antigen (EBNA), allows maintenance as a plasmid. The plasmid can be selected by a hygromycin resistance (*hyg*^{R}) gene expressed using a thymidine kinase promoter and polyadenylation signal, and sequences in the multiple cloning site can be expressed with the cytomegalovirus promoter (Pro CMV) and polyadenylation signal (PolyA). The plasmid can replicate as a shuttle vector in *E. coli*.

retrovirus from being packaged as well. This can be done by three methods (Figure 9.29): deletion of the packaging signal in the helper genome, deletion of multiple parts of the helper genome, or division of the helper genome into several pieces.

Other viruses used as a basis for vector construction include adenovirus, polyoma virus, which is closely related to SV40, both being members of the papovavirus family, Epstein–Barr virus and bovine papilloma virus (e.g. Figure 9.30).

In the last few years there have been significant advances in the development of MACs (or human artificial chromosomes), capable of carrying large stretches of DNA (in excess of 100 kb) and which can be maintained in the long term without the requirement for selection. A gene that is located within a large stretch of DNA may be more effectively regulated than a gene in an anomalous genomic environment, and this may be important for medical applications. A number of different approaches have been taken to the construction of human artificial chromosomes. In general, they involve either the 'trimming down' of an existing chromosome, or the assembly of a new chromosome from individual components. The most important of these components is the centromere, and in human artificial chromosomes this is usually constructed from tandemly repeated copies of a 171-bp alpha-satellite DNA sequence, found at the centromeres of primate chromosomes and containing recognition sequences for centromere-binding proteins. The generation of these artificial chromosomes is technically very difficult, and this is in part due to the need to manipulate very long stretches of DNA in vitro.

Promoters

CONSTITUTIVE PROMOTERS

Promoters used in any of the vectors described can be divided conveniently into a number of groups: those that are constitutive and those that are induced in response to various signals.

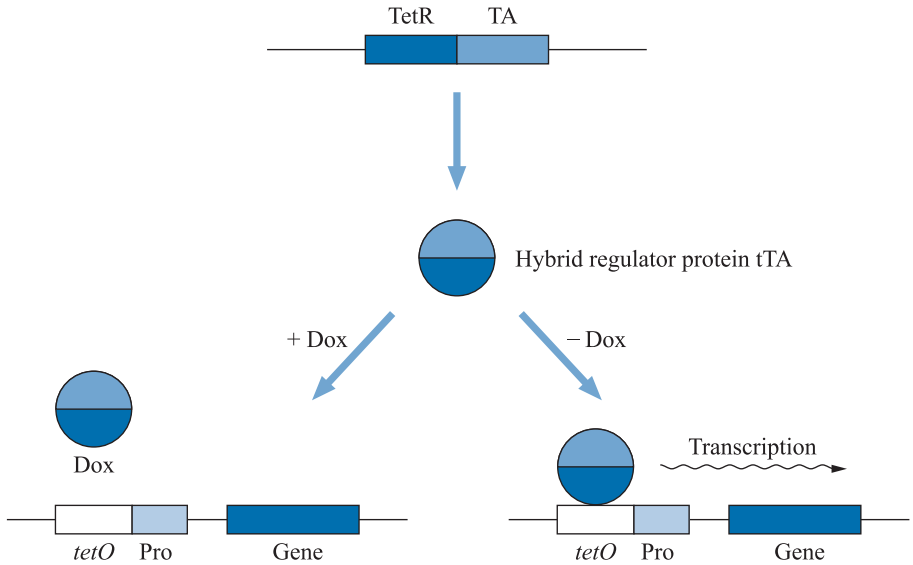


Fig 9.31 The Tet-off system. The hybrid regulator protein tTA comprises the tetracycline repressor domain (TetR) and a transcriptional activator domain from herpes simplex virus, TA. In the presence of doxycycline, the hybrid regulator is unable to bind at *tetO* and transcription from the promoter Pro is inactive. In the absence of doxycycline, the regulator binds and transcription is activated.

Constitutive promoters are frequently derived from viruses. They include the SV40 promoters, the Rous sarcoma virus promoter, the adenovirus major late promoter and the cytomegalovirus immediate early promoter. Hybrid promoters exist, containing elements from, for example, cytomegalovirus and human immunodeficency virus. Enhancers are often included, too, and frequently used ones are derived from SV40, Rous sarcoma virus and cytomegalovirus. However, some enhancers confer a significant degree of tissue specificity and may be unsuitable for general constitutive expression vectors. A number of vectors contain constitutive promoters derived from mammalian genes, such as elongation factor 1 alpha or ubiquitin, rather than viral ones.

TETRACYCLINE-REGULATED PROMOTERS

A very effective control system uses a mechanism for the control of gene expression in response to tetracycline (or analogues) that is derived from the bacterial transposon Tn10. In the naturally occurring bacterial system, a tetracycline-binding repressor protein, TetR, binds in the absence of tetracycline to a control sequence, *tetO*, and inactivates transcription. The method as it has been modified for artificial control of gene expression in mammalian cells (Figure 9.31) has two components. One is the TetR protein fused to part of a transcription-activator protein from herpes simplex virus. This is known as the transactivator protein, tTA. The second component is a hybrid promoter containing the prokaryotic *tetO* control sequence fused to a mammalian promoter (originally derived from cytomegalovirus) with low basal activity. In the absence of tetracycline (or an analogue, doxycycline) the tTA protein binds to the *tetO* control sequence and, because of the herpes simplex virus activation domain,

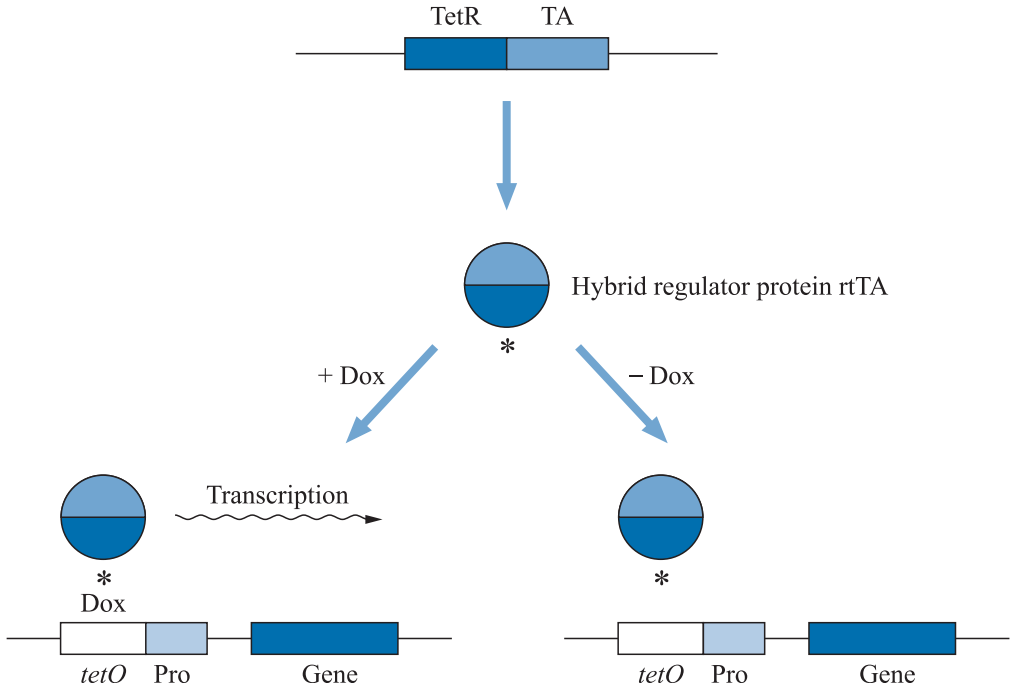


Fig 9.32 The Tet-on system. Abbreviations are as for Figure 9.31. In the Tet-on system, the hybrid regulator binds in the presence of doxycycline, activating transcription.

activates transcription from the mammalian promoter. (At first sight, this may seem confusing, because the original function of TetR is as a repressor! Here, it has been turned into an activator by fusion with the herpes simplex virus transcription activator protein.) The system, therefore, requires a host cell that has been engineered (i) to express the tTA protein and (ii) contains the gene to be expressed under the control of the hybrid promoter. Expression of the gene is inactivated by addition of the tetracycline analogue doxycycline. This system is particularly useful, as the control of expression is very tight – there is very little expression in the presence of doxycycline, so it is very good for expressing proteins that would otherwise be toxic. This is sometimes referred to as the Tet-off system.

The system has been modified by alteration of a few residues of the TetR part of the hybrid control protein (indicated by the asterisk in Figure 9.32). This modified protein is referred to as rtTA, as it has the reverse DNA-binding properties of tTA. It binds target DNA only in the *presence* of doxycycline, and not in its absence. The result of this is that a gene under the control of the hybrid promoter is usually inactive. If doxycycline is added, then it becomes active. This is sometimes referred to as the Tet-on system (Figure 9.32).

GLUCORTICOID-INDUCIBLE PROMOTERS

A widely used inducible promoter is the glucorticoid-responsive element from the long terminal repeat (LTR) of the mouse mammary tumour virus (MMTV). The LTR contains a glucorticoid-responsive element, to which a hormone-receptor complex can bind, activating transcription from the promoter. An example of a vector based on

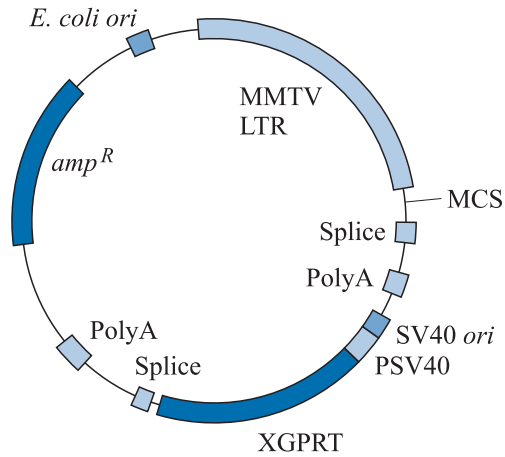


Fig 9.33 Plasmid pMSG (7.6 kb). This contains a mouse mammary tumour virus LTR sequence (MMTV LTR), including a glucocorticoid-inducible promoter. There are also SV40 splice and polyadenylation sites (Splice, PolyA). An SV40 origin—promoter region (SV40 *ori* and PSV40) drives expression of a bacterial XGPRT gene for use as a selectable marker in mammalian cells. There are also splice and polyadenylation sites to aid the XGPRT expression, and a prokaryotic origin and marker for propagation in *E. coli*.

this is pMSG (Figure 9.33), which includes elements from SV40 and pBR322, and a selectable marker (the *E. coli* xanthine-guanine phosphoribosyl transferase gene under the control of the SV40 early promoter). Induction of expression requires treatment with a suitable hormone (or analogue), such as dexamethasone, and the presence of the receptor in the cells. This can be provided either by using a cell type that is already hormone responsive, or by artificial expression in the cells of a receptor protein gene. This approach can be modified to improve expression further by incorporation of several copies of the glucocorticoid response element in the vector.

METAL-INDUCIBLE PROMOTERS

Another important group of promoters comes from the genes for metallothioneins. These are a group of proteins rich in cysteine residues, which chelate heavy metals to reduce their toxicity. The promoters can be induced by treatment of cells with heavy metal ions, such as zinc and cadmium. Inducibility depends on the presence of binding site for a control protein, and this can be incorporated into other promoters too. Use of metal-inducible promoters has two disadvantages: the heavy metal ions used may be damaging to the cells involved; and the control is not particularly tight, so that there may be significant levels of expression in the absence of the metal ions.

HEAT-SHOCK INDUCIBLE PROMOTERS

When cells are exposed to a heat shock, a specific set of genes is activated. This is mediated by the binding of a transcription factor to an element upstream from the promoters. This element can be used either with its normal promoter or in association with others to give controlled expression.

Other considerations

The amount of expression of DNA inserted into a vector does not depend only on the promoter used. If sequences are integrated, then

their chromosomal location is important. Flanking the inserted gene by A elements (or scaffold attachment regions), which are sites for attachment of the chromatin to the chromosomal scaffold, may help to reduce the position effects resulting from chromosomal location. A similar result may be obtained by the use of very long stretches of transforming DNA carried by artificial chromosomes. The presence of introns in the sequences to be expressed is also important, as are other factors such as poyadenylation of the transcript and efficient translation initiation.

9.10.2 Restricted areas of intact organisms

Introduction of DNA into a small region of tissue from an intact organism may be useful, e.g. in studies involving expression or in gene therapy (the introduction of DNA into an individual to treat a genetically determined disorder). For ethical reasons, it is generally considered that genetic modification of humans should not involve the germ line.

The principles for vector construction already outlined apply, although we are more likely to be dealing with a transient expression system. There are many ways of introducing DNA into target tissue, including microprojectile bombardment and direct injection. The latter is particularly suitable for some tissues, such as muscle, which forms a syncytium. The DNA may be complexed into liposomes with conventional transfection agents, and inhalation of liposomes has been used as a means of introducing DNA into the airways in an attempt to treat conditions such as cystic fibrosis. DNA introduction by viral infection is also widely used. For human gene therapy, the most commonly used viruses are retroviruses, adenovirus, adeno-associated virus and herpes simplex virus. There are concerns over the safety of viral vectors in humans, however.

9.10.3 Whole organisms

Generation of transgenic organisms

There are three main methods for generating whole organisms that are transgenic, summarized in Figure 9.34. One is to treat cleavage-stage embryos in vitro with recombinant retroviruses and then return them to the uterus of a foster mother. A more widely used approach is to isolate one-cell embryos (at which stage they are in the oviduct) and micro-inject DNA directly into the pronucleus. The embryos are then returned to a foster mother's oviduct, where development takes place. This approach usually results in the integration of DNA into the genome at random sites (which may also interfere with the function of endogenous genes), and sometimes multiple sites. A third approach exploits the ability to grow and manipulate mouse embryonic stem cells (ES cells) from the inner cell mass of a developing blastocyst. These pluripotent cells can be genetically modified in vitro and then reinjected into the blastocoel of a developing embryo. The injected embryo is then transferred to

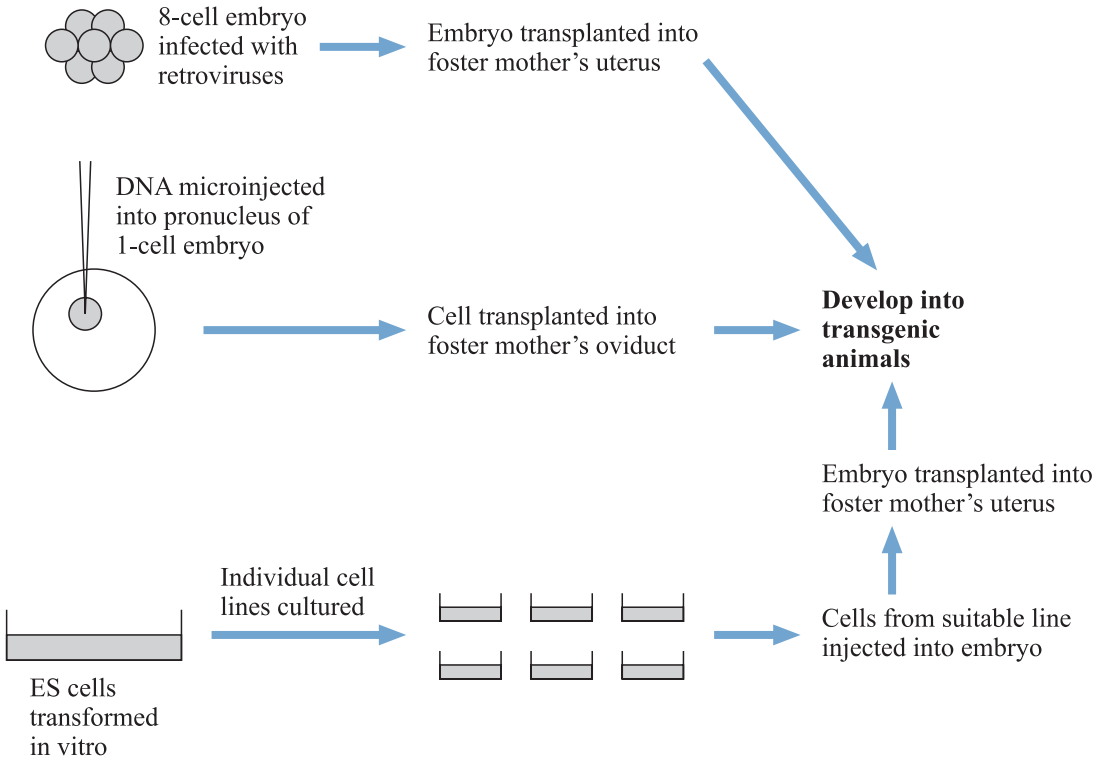


Fig 9.34 Three strategies for generating transgenic animals. See text for details.

a foster mother's uterus. The ES cells colonize the embryo during development, but the embryo does not become entirely composed of ES-derived cells, and the resulting mice will therefore be chimeric. Some ES-derived cells will be incorporated into the germ line. Matings with such mice can generate progeny that are transgenic and not chimeric, if a progeny animal is produced from a transgenic gamete. Producing transgenic mice in this way is analogous to the generation of transgenic *Drosophila* described above. Use of suitable genetic markers, such as coat colour, allows chimeric individuals to be identified easily. Very large stretches of DNA can be acquired, including artificial chromosomes.

The use of the three approaches, especially the last two, for the random integration of incoming DNA is relatively straightforward. To bring about expression of incoming sequences, the usual considerations about promoters, polyadenylation and so forth must be applied. The promoters described above can be used, as well as tissue-specific promoters. The Tet-on and Tet-off systems are particularly useful. If the rtTA transactivator protein from the Tet-on system is placed under the control of a tissue-specific promoter, then whichever gene is placed under the control of the rtTA-regulated promoter will be activated by the addition of doxycycline, but only in certain tissues (the ones in which the transactivator is expressed). This approach has been exploited to allow the selective destruction of certain tissues to study the effects that result. In this strategy, the rtTA transactivator protein is placed under the control of an

appropriate tissue-specific promoter, and in turn the transactivator controls the expression of an intracellular toxin. In the absence of doxycycline, the toxin is not expressed. When doxycycline is added, the transactivator binds to its promoter, leading to expression of the toxin from the doxycycline-responsive promoter, but only in those cells that were expressing the transactivator. Those cells die. The technique is sometimes referred to as ablation.

Gene targeting

If it is simply sufficient that a sequence be expressed in a transgenic organism, then it will not be necessary to ensure integration of the DNA at the normal chromosomal location. For some purposes, though, it may be necessary to ensure such integration, perhaps to improve the regulation of expression of integrated DNA or to disrupt an endogenous gene. Obtaining integration at a particular site is often called gene targeting, and it can be achieved using ES cells. Targeting relies on the fact that, in ES cells, homologous recombination can take place between incoming DNA and the chromosome at the site of the corresponding sequences. The principle is that ES cells are screened after transformation for the 'correct' homologous recombination, and only those that have undergone a suitable integration event are injected into a developing embryo. Mice in which a gene has been inactivated through gene targeting are called knock-out mice. Replacing a gene with a different one is sometimes called knocking in.

Integration events may be insertional, involving a single crossover between the incoming DNA and the target gene, or replacement, involving two crossovers. Selection for acquisition of the incoming DNA can be achieved quite easily, e.g. by selection for a marker such as resistance to neomycin or G418 (Figure 9.35). However, the difficulty is that (although homologous recombination can take place) non-homologous recombination can also take place, causing sequences to be integrated at other sites in the genome. Further screening, therefore, has to be done after the initial neomycin selection for transformation, in order to identify those cells where homologous insertion has taken place. It is useful if this screening can be done selectively, as the frequency of homologous recombination may be low compared with random integration. A typical example, involving two different rounds of selection, is summarized in Figure 9.36. The incoming DNA contains two markers. Selection of cells that have acquired the first (in this case resistance to G418) identifies those that have integrated DNA. The second marker is placed so that it is lost when integration is homologous, but is likely to be retained when integration is not homologous (as incorporation of molecules by non-homologous recombination usually happens from the free ends). The second round of screening requires selection for cells that have *not* acquired the second marker, which in this case is the herpes simplex virus TK gene. Expression of this renders cells sensitive to the nucleoside analogue gancyclovir (see above), so cells

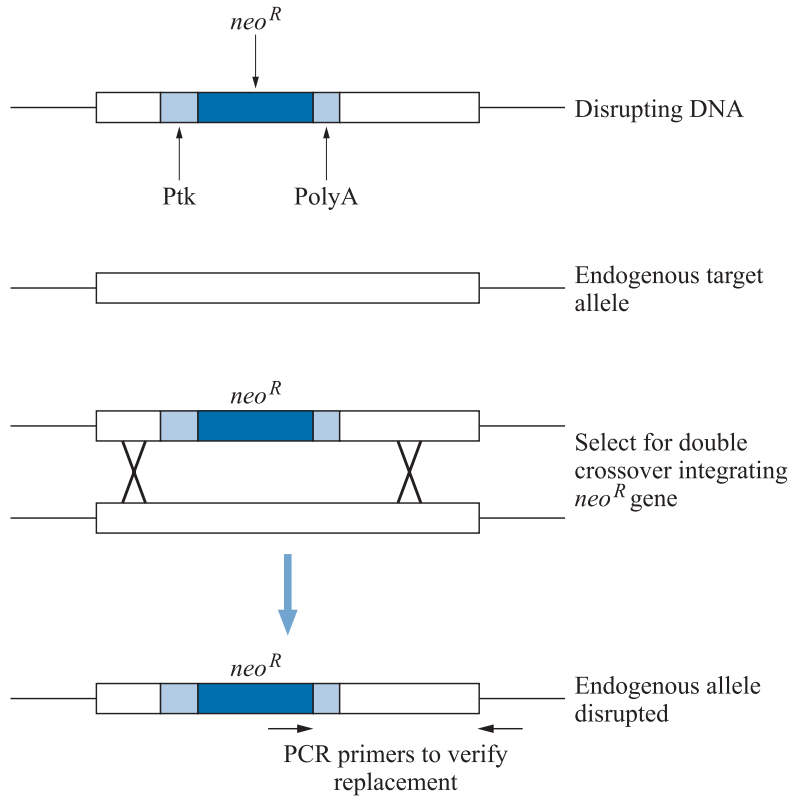


Fig 9.35 Gene disruption in an ES cell. The disrupting DNA contains a gene (*neo*^{R}) conferring resistance to neomycin or the antibiotic G418 (with a thymidine kinase promoter (Ptk) and a polyadenylation site) replacing part of the target gene. Integration of the *neo*^{R} gene at the chromosomal site of the target gene can be demonstrated by the formation of a correctly sized PCR product using appropriate primers.

that are resistant to both G418 and gancyclovir are likely to have had a targeted disruption.

Whether or not the selection against random insertion is used, it is necessary to analyse the genome of the transgenic cells to confirm that the site of integration is correct. This was done by Southern blotting in early experiments, using DNA from cultures derived from individual transformed cells. Screening many cell lines this way is laborious, and this approach has been superseded by PCR analysis. In the latter case, PCR is carried out using one primer that corresponds to a sequence within the incoming DNA and one that corresponds to a sequence close to the intended site of insertion (Figure 9.35). Homologous integration of sequences would give rise to fragments of predictable size when Southern blots were probed, or likewise, PCR products of predictable size. The cells where homologous recombination has taken place can then be taken on to generate transgenic organisms.

Sometimes it is desirable to make minor changes to a target gene, without incorporating a selectable marker at that site, in case the marker disturbs expression. One approach relies on the fact that cells that have incorporated DNA at one site are also more likely to have incorporated unlinked sequences at a different site (cotransformation). Cells are co-transformed with DNA bearing the mutated form of the target gene and with separate molecules bearing

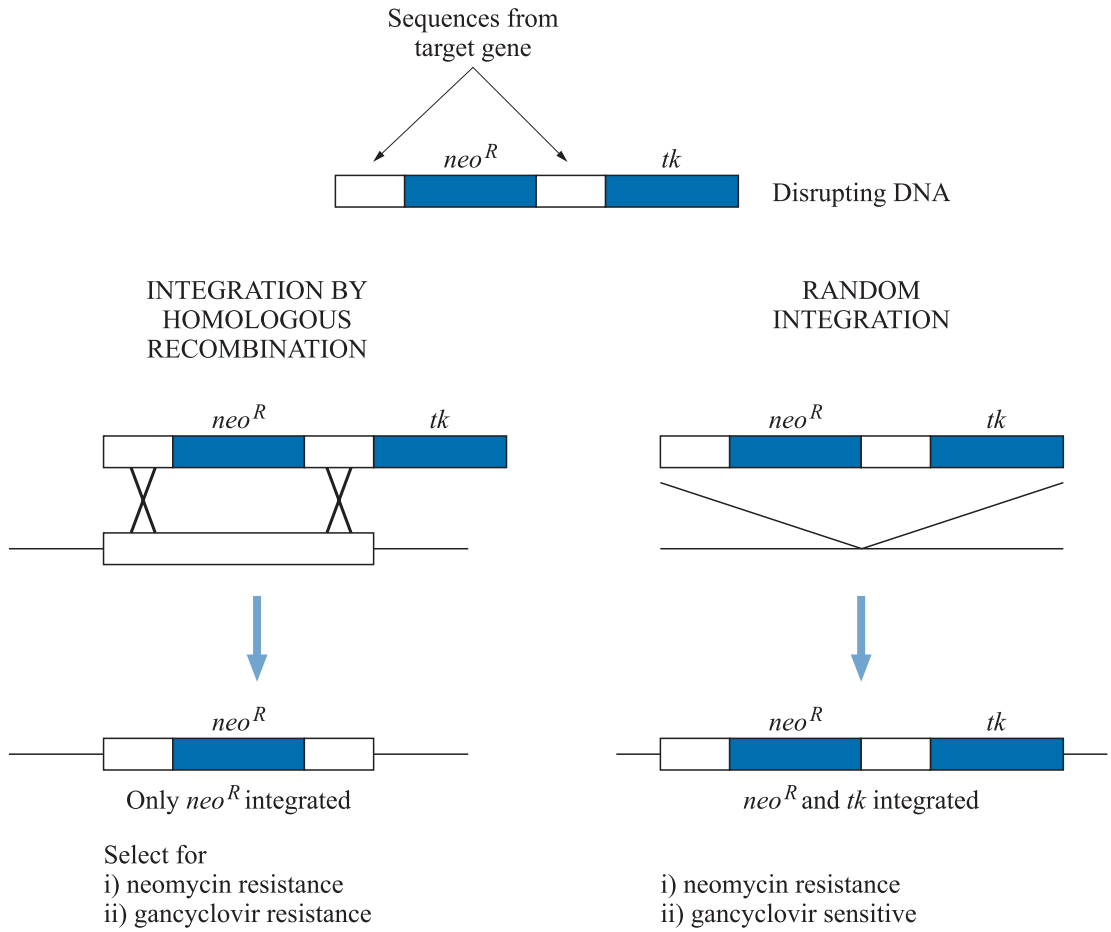


Fig 9.36 Selection for homologous recombination in gene disruption. Incoming disrupting DNA carries the *neo*^{R} gene (see Figure 9.35) and thymidine kinase (*tk*) markers. Homologous recombination between the chromosome and the sequences flanking the *neo*^{R} gene does not integrate the thymidine kinase gene; random integration (stimulated by the free end of the molecule) does.

a selectable marker. Cells that have acquired the marker by random integration are selected and then screened (e.g. by PCR) for co-acquisition of the mutation by homologous recombination at the target site.

The gene targeting approach ultimately generates individuals that are heterozygous for inactivation of the gene in question. It will often be necessary to generate homozygotes, which can be done by crossing heterozygous individuals and screening the progeny for homozygosity. (If the inactivation is a recessive lethal trait, then homozygotes will not be recovered. An elevated level of embryo mortality, resulting from the death of the homozygotes, may be observed.) Alternatively, growth of heterozygous cells in the presence of elevated concentrations of the selective agent (such as G418) can sometimes spontaneously generate homozygotes, which are presumably generated from the heterozygotes by gene conversion or a related process.

Conditional gene targeting

Conditional gene targeting can be used to bring about the controlled inactivation of a gene, and avoid some of the consequences when inactivation of a target gene is deleterious. Transgenic organisms are constructed in which the gene of interest is replaced by a copy flanked by *loxP* sites from the Cre-*lox* site-specific recombination system (see Chapter 7). The gene for the Cre recombinase is placed under a controllable promoter. Induction of this promoter leads to expression of recombinase and excision of the gene located between the *loxP* sites.

Applications

The applications of transgenic animal technology are clearly extremely broad. For example, for studies on gene regulation, transgenic animals provide a more realistic system than does the use of cultured cells. The genetic modification of economically important animals may be financially useful, either for the effects on the animals themselves or in the production of pharmaceutically important polypeptides.